UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Education Data Mining Para Apoio à Gestão Estratégica da Identificação de Perfis Evasivos e Atenuação da Evasão Escolar no Ensino Superior

Kelly Joany de Oliveira Santos

São Cristóvão – Sergipe

2020

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Kelly Joany de Oliveira Santos

Education Data Mining Para Apoio à Gestão Estratégica da Identificação de Perfis Evasivos e Atenuação da Evasão Escolar no Ensino Superior

Proposta de Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Dr. André Britto de Carvalho

São Cristóvão – Sergipe

2020

Resumo

A mineração de dados educacionais é um campo de pesquisa que visa extrair informações de grandes conjuntos de dados de cunho didático. É uma área multidisciplinar com diversas possibilidades que ganha cada vez mais destaque em unidades de ensino. Esta tecnologia recente possibilita por meio de técnicas de mineração de dados gerar informação a partir de atributos como, por exemplo, dados de desempenho acadêmico. Pesquisas recentes apontam que estas técnicas podem identificar alunos que apresentam indicadores para o abandono aos estudos. Sendo assim, a evasão escolar é um dos principais desafios das universidades e demais organizações que buscam compreender os motivos que levam o discente a evadir do curso escolhido. No entanto, modelos e perfis de alunos evasivos ainda são pouco estudados, o que leva a falta de consolidação sobre as razões que elevam a evasão. Esta pesquisa propõe uma abordagem para a identificação de perfis de estudantes evasivos no âmbito de unidades de ensino, com o intuito de apoiar decisões que atenuem a evasão escolar no ensino superior. A metodologia apresentada consiste em obter e criar a base de dados para os estudos, a realização de análise preliminar do ambiente observado e o método que avalia o desempenho de alunos em disciplinas. Para avaliar a abordagem proposta, este trabalho realizou um estudo de caso com os dados de alunos da Universidade Federal de Sergipe, em cursos do Departamento de Computação, para a aplicação específica do problema, consolidados em um *Data Warehouse*, que permitiu investigar a evasão entre os anos de 2007 a 2018. Nesta pesquisa, são apresentados problemas comuns enquanto utilização de mineração de dados educacionais, como: a seleção de atributos, estruturação dos dados, valores errôneos e correções dos mesmos. O resultado inicial apresenta a acurácia de 98,647% para o algoritmo árvore de decisão, o que permite concluir que este é o algoritmo mais indicado para a tomada de decisão em cenários onde se busca a atenuação da evasão escolar.

Palavras-chave: Mineração de Dados Educacionais, *Business Inteligence*, Educação.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CC	Ciência da Computação
DCOMP	Departamento de Computação
DECAT	Departamentode Estatística e Ciências Atuarias
DED	Departamento de Educação
DFI	Departamento de Física
DM	*Data Mart*
DM	*Data Mining*
DMA	Dificuldade Média do Aluno
DMA	Departamento de Matemática
DMP	Dificuldade Média do Período
DMT	Dificuldade Média da Turma
DPS	Departamento de Psicologia
DS	*Data Sources*
DW	*Data Warehouse*
EC	Engenharia da Computação
EDM	*Education Data Mining*
ETL	*Extract, Trasnform and Load*
SGBD	Sistema Gerenciador de Banco de Dados
SI	Sistemas de Informação
OLAP	*Online Analytical Processing*

Sumário

1 Introdução

De acordo com Censo da Educação Superior 2015, 11% dos alunos que entraram na graduação em 2010 desistiram já no primeiro ano. Até 2014, quase metade (49%) dos estudantes saíram dos cursos que haviam optado em 2010 (TEIXEIRA, 2015). Sendo assim, a evasão é um dos grandes problemas que afligem as instituições de ensino em geral. Segundo FILHO (2007), diversas situações podem ser caracterizadas como evasão, tais como: o trancamento de um curso por um estudante, a desistência por falta de interesse, a falta de recursos financeiros do aluno, motivo de doença, gravidez precoce, a desistência devido à incompatibilidade de horários das aulas com o mercado de trabalho ou ainda quando o estudante dá início a carreira profissional. A busca de suas causas tem sido objeto de muitos trabalhos e pesquisas educacionais, como por exemplo (FILHO, 2007), (SANTOS, 2017), (MANHAES et al., 2012).

Ambiel (2015), após revisão da literatura, elenca os principais fatores que estão relacionados ou podem influenciar em cenários de evasão escolar. Dentre as razões podemos destacar a qualidade de ensino insuficiente, ou seja, a baixa qualidade de ensino antes de entrar na graduação; as relações sociais insatisfatórias em relação a outros alunos, docentes e colaboradores da instituição; a falta de opções em atividades extracurriculares; a dependência financeira para custear os estudos ou a necessidade de trabalho para complemento de renda; características familiares sociodemográficas; direcionamento errôneo para a escolha do curso; e o rápido ingresso na graduação logo após o término do ensino médio. Ainda em seu estudo o autor construiu, a partir de 81 itens, motivos que poderiam induzir ou sugestionar a decisão do aluno a evadir do curso superior escolhido.

A escala de motivos para evasão do ensino superior, obteve de maneira geral o agrupamento de quatro razões distintas: motivos institucionais, estes se relacionam diretamente com a qualidade do corpo docente e apoio aos alunos, infraestrutura da instituição e a falta de serviços direcionados ao estudante; motivos pessoais, estes englobam as incertezas em relação a estar no curso certo e influência familiar; motivos relacionados a dificuldades financeiras e/ou dificuldade

em conciliação dos estudos e a jornada de trabalho; motivos relacionados às incertezas de carreira, tanto em realização pessoal dos discentes quanto em aspectos do mercado de trabalho.

No Brasil diversas organizações governamentais buscam por decisões estratégicas, visando o controle da taxa de evasão (FILHO, 2007). Nessa direção, o Documento Orientador para a Superação da Evasão e Retenção na Rede Federal de Educação Profissional, Científica e Tecnológica sugere que cada instituição da Rede Federal elabore e desenvolva um Plano Estratégico de Intervenção e Monitoramento para Superação da Evasão e Retenção (MEC, 2014).

Seguindo este direcionamento, já é possível encontrar contribuições sólidas que apontam o interesse dos pesquisadores brasileiros nesta área (BALANIUK et al., 2011), (MARTINS et al., 2017), (MANHÃES; CRUZ; ZIMBRÃO, 2014). Deste modo é possível analisar a evasão sob um novo ponto de vista, o que faz a mineração de dados ganhar cada vez mais relevância em estudos e pesquisas científicas aplicadas em cenários de ensino.

A mineração de dados é considerada uma opção para pesquisa quando se deseja extrair informações de forma eficiente de um grande conjunto de dados. Este conceito surgiu em meados dos anos 1990 com uma nova abordagem para análise de dados e descoberta de informações. A primeira Conferência da ACM em que foi abordado este tema foi no evento *Conference on Knowledge Discovery and Data Mining* nos EUA em 1995, no entanto, o termo mineração de dados teve seu primeiro registro muito posteriormente, em 2009. Apesar da mineração de dados se originar a partir de trabalhos realizados no campo da estatística, este foi um divisor de águas que permitiu o reconhecimento de padrões, novos designs para banco de dados, visualização de informações, trabalhos relacionados na área da inteligência artificial, aprendizagem de máquina e outras possibilidades (YOO et al., 2011).

A mineração de dados educacionais, *Education Data Mining* (EDM) é considerada uma área de pesquisa multidisciplinar que reúne pesquisadores da ciência da computação, educação, psicologia e estatística. A revisão de literatura fornece amplo conhecimento sobre como o EDM evoluiu desde o seu início até o cenário atual. É evidente que com os avanços nas técnicas de mineração de dados os mesmos foram adaptados ao EDM. Esta, por sua vez, avançou como ferramenta para melhorar a gestão e administração de sistemas educacionais (PUARUNGROJ et al., 2018).

Vários estudos lidam com a implementação de técnicas de mineração de dados para detectar alunos com alto risco de evasão em instituições de ensino. Para que este objetivo seja alcançado os algoritmos utilizam dados pessoais, demográficos, conhecimento educacional, e também o progresso na unidade de ensino. Esses estudos demonstraram que o sucesso acadêmico depende em grande parte destes fatores (MáRQUEZ et al., 2016), (AL-SHABANDAR et al., 2017), (BOGARÍN et al., 2014).

Neste contexto, nos últimos anos a possibilidade de prever o abandono dos alunos dos cursos de ensino superior tornou-se uma pesquisa importante e desafiadora para as universidades

(KOSTOPOULOS; KOTSIANTIS; PINTELAS, 2015). A conclusão de um curso superior é de grande importância especialmente para as entidades que oferecem cursos a distância, uma vez que o abandono e a retenção de alunos podem ser utilizados como indicadores importantes da eficácia da aprendizagem e qualidade de ensino. Além destes pontos, a identificação precoce de estudantes com alto risco de evasão possibilita a equipe acadêmica apoiar os discentes com baixo desempenho através de programas de intervenção, *workshops* e materiais de aprendizado diferenciado (LIANG; LI; ZHENG, 2016).

A mineração de dados é uma das técnicas que compõem o *Business Intelligence* (BI), que pode ser apresentado como uma tecnologia, arquitetura; além de ferramenta ou sistema que coleta e armazena dados. Através do BI é possível explorar os dados utilizando métricas analíticas, proporcionar a criação de relatórios e consultas, e entregar informação com a finalidade de auxiliar a tomada de decisão de instituições (KHAN; QUADRI, 2019). Sistemas educacionais podem produzir uma massa expressiva de dados e ferramentas BI podem auxiliar as visualizações deste grande volume de informações através de *reports*. Deste modo, a utilização de BI dentro de cenários de ensino superior pode otimizar o processo de tomada de decisão acadêmico relacionado aos discentes ainda não formados.

1.1 Problemática

Presente no sistema educacional brasileiro nos seus diversos níveis e modalidades, a evasão escolar tem atingido desde a educação básica até a superior, gerando prejuízos sociais, econômicos, políticos, acadêmicos e financeiros a todos os envolvidos no processo educacional, desde o estudante até os órgãos governamentais. Diagnosticar suas causas, compreender como esse processo ocorre nas instituições de ensino e conhecer a visão dos gestores educacionais sobre esta problemática pode auxiliar na compreensão deste fenômeno social e nas ações preventivas para sua redução (SANTOS, 2017).

De acordo com Lobo (2012), reduzir a evasão escolar custa até seis vezes menos que custear um novo estudante até a instituição de ensino. Deste modo, mitigar a evasão estudantil é a maneira mais eficiente de aumentar o número de matrículas. Incentivar que o aluno conclua seu curso, com qualidade, significa evidenciar que a instituição, seja pública ou privada, atingiu o objetivo instituído.

Além da evasão, as informações inadequadas são uma das principais causas de decisões equivocadas. A ausência de metodologia contribui diretamente para a má interpretação dos dados disponíveis. Lima, Júnior e Nascimento (2017) evidenciaram que 72% das empresas brasileiras entrevistadas não utilizavam um método específico para o desenvolvimento de aplicações de BI alinhado ao planejamento estratégico da organização. Portanto, torna-se fundamental a utilização de sistemas aliados aos objetivos da entidade.

Nesta conjuntura, um número de pesquisas foram beneficiadas através da utilização de

técnicas de mineração de dados na área de educação. Além da predição de desempenho dos alunos buscou-se compreender os fatores que levam o estudante a evadir da instituição de ensino. (MANHAES et al., 2012), (FILHO, 2007), Marquez-Vera, Morales e Soto (2013).

1.2 Objetivos

1.2.1 Objetivo Geral

Este trabalho tem como principal propósito realizar, através de algoritmos de mineração de dados educacionais, um estudo de caso que visa entender as razões que levam os estudantes a evadirem das unidades de ensino. Se busca diretamente através da criação de novos atributos propor uma abordagem robusta, em relação aos demais trabalhos relacionados, que permita efetivamente analisar os discentes através da disciplina a ser cursada. Como estudo de caso para o contexto da evasão escolar foi selecionado o âmbito da Universidade Federal de Sergipe cujo os dados dos cursos de Ciência da Computação, Sistemas de Informação e Engenharia da Computação foram considerados para a aplicação específica do problema.

1.2.2 Objetivos Específicos

Para realizar o objetivo geral, podemos destacar os seguintes objetivos específicos:

• Realizar mapeamento sistemático da literatura, com o intuito de identificar quais são os principais algoritmos de mineração de dados encontrados nos trabalhos de EDM em contextos de evasão escolar;

• Realizar a análise experimental preliminar com os algoritmos selecionados de acordo com o mapeamento sistemático, para os cursos de Ciência da Computação, Sistemas de Informação e Engenharia da Computação desde o 1º ao 6º semestre da Universidade Federal de Sergipe - UFS.

• Propor abordagem correspondente de acordo com os resultados encontrados.

1.3 Metodologia

A metodologia utilizada para o trabalho consiste, em termos de classificação, de uma pesquisa exploratória e descritiva. A pesquisa exploratória está representada através de um mapeamento sistemático, no qual, foram selecionados as técnicas de mineração de dados que serão utilizados neste estudo. A pesquisa descritiva está relacionada diretamente ao estudo de caso, onde é realizado a interpretação dos resultados encontrados com as questões de pesquisa propostas. Além destes pontos também busca-se oportunidades de replicação do modelo de perfil evasivo proposto.

1.3.1 Obtenção e criação da base de dados para estudos experimentais

Para que seja possível interpretar corretamente as informações presentes neste grande volume de dados é necessário uma organização que reflita as grandezas aqui representadas de maneira acertiva. Neste momento foram adotados os seguintes passos:

• A obtenção de dados brutos gerados a partir da plataforma acadêmica SIGAA oriunda da Universidade Federal de Sergipe;

• A estruturação dos dados de modo a ser possível um relacionamento entre os mesmos;

• Aplicar a tecnologia ETL de modo a tratar os dados obtidos;

• Espelhar a modelagem definida em um sistema gerenciador de banco de dados - SGBD;

• Otimizar os resultados encontrados seja através de ETL ou outras técnicas de limpeza dos dados.

• Possibilitar a visualização dos resultados obtidos em *reports* através da ferramenta BI open-source Pentaho.

A partir desta etapa é possível realizar estudos experimentais, além de facilitar a exploração dos dados através de cruzamentos antes inviáveis por não existir relacionamentos entre as entidades e estrutura física para tal abordagem. A Seção 3 detalhará os dados brutos e a estrutura adotada para os estudos iniciais.

1.3.2 Análise preliminar

A análise preliminar tem como objetivo avaliar a base de dados citada na Seção 1.3.1, além de determinar qual algoritmo será utilizado no decorrer desta pesquisa. Foram considerados os períodos de 2010.1 a 2018.1 para a contextualização específica do problema em cursos relacionados a Computação. Nesta etapa será avaliado o desempenho de algoritmos que foram elencados através do mapeamento sistemático, qual destes é o mais adequado a este cenário e a visualização destes resultados em *reports* através da ferramenta BI open-source Pentaho. A Seção 4 detalhará o desenvolvimento desta etapa, bem como elencará as considerações encontradas após a análise preliminar.

1.3.3 Abordagem para avaliar o desempenho de alunos em disciplinas

A abordagem para avaliar o desempenho de alunos em disciplinas busca apresentar uma análise aprofundada do contexto evolutivo do aluno por disciplina realizada através das técnicas EDM. É proposto uma abordagem computacional orientada a cinco perguntas que foram formuladas para refletir os principais problemas da instituição analisada. Nesta abordagem é realizada uma análise micro da situação em que o aluno se encontra em relação a disciplina matriculada. Para que esta análise seja possível, novos atributos foram criados para mensurar

dimensões como, por exemplo: o histórico da disciplina a ser cursada, o histórico do aluno em si e a grade curricular referente ao curso superior escolhido. A Seção 5 detalhará o desenvolvimento desta etapa além de mostrar como responder a cada questão elaborada.

1.3.4 Gestão Estratégica e a Abordagem para avaliar o desempenho de alunos em disciplinas

O conhecimento gerado através desta pesquisa pode auxiliar a tomada de decisão por parte de docentes e gestores acadêmicos, além de gerar valor para as próximas turmas do Dcomp. A partir da análise preliminar e das questões de pesquisa abordadas, em nosso estudo de caso, é possível identificar pontos de melhoria no processo de ensino atual. As informações coletadas de forma macro podem ser utilizadas dentro de projetos internos para direcionar os esforços para a retenção de alunos, e em relação ao estudo de caso apresentado, no Dcomp.

Além destas informações macros, a ferramenta Knime que será apresentada na Seção 2.1.7.1, traz a visão/seleção de turmas e disciplinas disponíveis, onde por sua vez, possibilita realizar o acompanhamento dos alunos matriculados. Assim sendo, através do modelo treinado que será relatado na Seção 5.4.1 é possível predizer se o aluno apresentará desistência/reprovação na turma matriculada, deste modo, o docente pode aplicar um ensino direcionado aos estudantes que já apresentam indícios a reprovação/evasão da disciplina.

1.4 Contribuições

Esta pesquisa além de proporcionar aos docentes e gestores acadêmicos uma nova visão/abordagem que pode ser adotada para aumentar a retensão de alunos em disciplinas, apresenta através da abordagem computacional, como a transformações de dados e a criação de novos atributos podem aumentar o desempenho do algoritmo e a reduzir a maldição da dimensionalidade em estudos voltados ao *Education Data Mining*.

A criação de novos atributos permite explorar novas possibilidades e otimizar o potencial que o algoritmo selecionado pode alcançar. Deste modo, esta abordagem pode fomentar outras formas de exploração na área de *Education Data Mining*, bem como, proporcionar outras visões dentro de pesquisas atuais.

1.5 Estrutura do Trabalho

O Capítulo 2 é dedicado ao referencial teórico, apresentando os principais conceitos que envolvem as técnicas mineração de dados utilizadas na realização deste estudo. Deste modo, também é apresentado a análise de trabalhos correlatos.

O Capítulo 3 apresenta o processo da criação da base de dados para os estudos experimentais. É apresentado a estrutura, os atributos e a modelagem adotada para esta pesquisa dentro do cenário estudado.

O Capítulo 4 apresenta um estudo preliminar da evasão bem como a utilização de algoritmos selecionados para a análise inicial do problema. Neste momento também é apresentado o desenvolvimento desta fase, os resultados encontrados e as considerações desta etapa.

O Capítulo 5 apresenta a abordagem para avaliar o desempenho de alunos em disciplinas. É apresentado em paralelo a criação dos atributos utilizados para realizar as inferências com a finalidade de auxiliar o algoritmo selecionado para a tomada de decisão no ambiente educacional.

O Capítulo 6 apresenta os resultados da abordagem computacional a partir dos experimentos realizados. Além dos resultados, também se busca responder as perguntas elaboradas sob o contexto desta abordagem.

Finalizando este trabalho, o Capítulo 7 apresenta as considerações finais onde são apresentadas as conclusões, bem como, as possíveis contribuições desta pesquisa e suas limitações. Ao final deste, são apontadas perspectivas para a continuidade desta pesquisa em trabalhos futuros.

2

Fundamentação Teórica

Esta Seção tem como principal objetivo contextualizar os conceitos básicos da área bem como, apresentar a discussão dos trabalhos relacionados que fomentaram este estudo. Em seguida é apresentado de maneira sucinta os resultados encontrados através do mapeamento sistemático elaborado na etapa inicial desta pesquisa.

2.1 *Background* Técnico

2.1.1 *Business Intelligence (BI)*

Para Lima (2017), o problema das organizações, atualmente, é que as empresas não conseguem declarar os objetivos estratégicos de forma explícita ou suficientemente clara, para que se possa verificar se tais objetivos têm realmente alcançado as metas e estão alinhados à TI. Este desafio não é ter a área de TI como um suporte, mas sim como parte de uma plataforma de negócio, servindo como elemento essencial à estratégia de negócio. Essa nova visão deve vincular o alinhamento estratégico da TI ao negócio da organização. Estes dois elementos precisam relacionar-se entre si, em busca da melhoria contínua e do sucesso da organização.

Business Intelligence – BI pode ser definido como um conjunto de tecnologias, processos, metodologias e arquiteturas que possibilitam a transformação de dados não trabalhados em informações, que por sua vez, são utilizadas para a tomada de decisões estratégicas, táticas e operacionais mais efetivas e eficientes (HANS; MNKANDLA, 2013). Para GARTNER (2018), BI pode ser compreendido como o processo que transforma dados em informação e, por meio da descoberta, possibilitar a transformação da informação em conhecimento.

BI pode ter uma significância diferente de acordo com o contexto a ser aplicado (ABREU, 2008). Sob uma visão tecnológica, o conceito de BI está diretamente relacionado ao processo de extrair, transformar, gerir e analisar os dados propostos, com o intuito de auxiliar a tomada

de decisão de gestores. Este procedimento é baseado, principalmente, em análise de grandes volumes de dados, com o objetivo de transferir o conhecimento em toda organização, do nível estratégico ao nível tático e operacional.

O *Business Intelligence*, pode ser através dos seguintes componentes: Data Sources (DS), *Extract, Trasnform and Load* (ETL), *Data Warehouse* (DW), o *Data Mart* (DM), *Data Mining* e ferramentas *Online Analytic Processing* – OLAP. Na figura 1, é exibida a relação entre os componentes.

Figura 1 – Tecnologias que compõem uma aplicação de *Business Intelligence* – BI.

Fonte: (LIMA, 2017)

O BI é visto como uma referência para diferentes tecnologias, como elencado na figura 1, deste modo é importante entender o conceito sobre cada ponto deste (CEDERBERG, 2010).

• *Data Sources* (DS) ou Fonte de dados - É a origem dos dados a serem trabalhados. Diversas fontes de dados são representadas como um desafio para a interpretação na visão dos clientes. A variedade de dados dentro de uma instituição torna-se, em alguns casos, não confiável.

• *Extract, Trasnform and Load* (ETL) - A ferramenta ETL é usada para extrair dados de transações diversas e realizar a coleta que será armazenada em DW.

• *Data Warehouse* (DW) - Coleta e armazena dados de diferentes fontes em um único lugar. DW pode ser definido como "um armazém de dados simples, completo e consistente obtidos de uma variedade de fontes e disponibilizado aos usuários de uma forma que eles possam entender e usá-los em um contexto de negócios"(DEVLIN, 1997). DW também contempla a estrutura de dados multidimensional.

• *Data Mart* (DM) - Pode se definida como a camada de acesso *frontend* do DW. Tem como um de seus objetivos ajudar a obtenção de dados para os usuários. Os *data warehouses* e *data marts* são necessários para utilização conjunta, uma vez que, a informação no DW não está organizada de modo a ser intuitivo a busca pela informação que é necessária.

• *Online Analytical Processing* (OLAP) - Possibilita a criação mais rápida de novos relatórios, que interpretam os dados e auxiliam a tomada de decisões.

• *Report* (Relatório) - É um documento com formato unificado que pode apresentar

análise estatística além da visualização dos dados tratados.

• *Data Mining* (DM) - É uma funcionalidade que agrega e organiza dados diversificados. Possibilita encontrar padrões preditivos, realizar associações, mudanças ou anomalias relevantes para as tomadas de decisões estratégicas.

Cada componente cumpre uma função fundamental no processo de BI, permitindo que, ao final deste processo, sejam entregues aos usuários da instituição as informações estratégicas ou a inteligência necessária para a tomada de decisão.

A implementação ou implantação de aplicações de BI devem incorporar aspectos de negócio, além dos aspectos técnicos. Isso significa que o desenvolvimento de aplicações de *Business Intelligence* devem ser realizados alinhados diretamente às estratégias de negócio da entidade (CEDERBERG, 2010). Segundo Hans e Mnkandla (2013), toda decisão oriunda das ferramentas BI devem ser baseadas no conhecimento sobre seu ambiente de negócio, pois os gestores dependem das informações valiosas disponibilizadas pelas ferramentas de BI.

2.1.2 *Business Intelligence (BI)* e visão estratégica

Um dos papéis fundamentais da tecnologia da informação é ser apoio para processos estratégicos. Os avanços tecnológicos, por sua vez, redefinem como as empresas e demais organizações realizarão as tomadas de decisão. A adoção de tecnologias permite a redução significativa de custos e a melhora de desempenho através de tarefas automatizadas (MARCH; HEVNER, 2007).

A utilização de BI pelas organizações é visto como predição de acontecimentos, por meio da predisposição demonstrada pelas informações existentes. Algumas organizações e entidades fracassam na manutenção da qualidade das informações que são provindas dos sistemas transacionais, que sustentam o sistema de BI, predispondo a situações em que não é possível perceber que as decisões estão sendo prejudicadas por dados errados (ISIK; JONES; SIDOROVA, 2012). Dentro deste contexto uma opção de ferramenta de BI a ser utilizada é o *Data analytics*. Este consiste basicamente na aplicação de tratamento estatístico em dados coletados, com o intuito de gerar predições e em dar sentido a esses dados, transformando-os em informação que ajudam a tomada de decisões e planejamento estratégico das empresas (CHEN; CHIANG; STOREY, 2012). Para Harriott (2013) *Analytics* se relaciona com o “como”, ou seja, como desenvolver e prover a informação de forma a realmente resolver os desafios da organização.

Construir as decisões futuras de uma organização em detrimento de informações equivocadas poderá induzir ao fracasso dos objetivos pré-definidos da instituição ou direcionar à extinção completa da organização. Para definir os caminhos a serem seguidos se demandam o conhecimento e experiência de todas as partes envolvidas, de modo a identificar quais os principais elementos são de fato importantes para o sucesso da empresa. Isto exige um planejamento de ações que devem garantir que quaisquer mudanças nos objetivos estratégicos possam ser

refletidas nas aplicações e demais sistemas de tomada de decisão. Essa vinculação de sistemas e estratégia é importante, de modo que todo esforço adotado no desenvolvimento da aplicação possa contribuir com os objetivos estratégicos (LIMA; JÚNIOR; NASCIMENTO, 2017).

2.1.3 *Data Mining*

Para Turban et al. (2013), data mining (DM) é um termo utilizado para descrever a descoberta (ou mineração) de conhecimento a partir de grandes quantidades de dados. É utilizado para identificar informações úteis através de padrões que podem ser apresentados como tendências, perfis, regras de negócio, modelos preditivos ou correlações. Esses modelos e padrões podem ser utilizados para guiar o processo decisório, bem como prever o efeito dessas escolhas (LAUDON; LAUDON, 2007).

Em um contexto de mineração de dados temos as atividades e as técnicas que nos direcionam ao objetivo que se desejam atingir. Deste modo é possível a realização da abordagem supervisionada e a não-supervisionada. Os métodos supervisionados necessitam de um conjunto de dados que possuem uma variável alvo principal pré-definida e os registros são categorizados em relação a ela, ou seja aplica-se a regra programada com precisão. Nos métodos não-supervisionados não há a necessidade de uma pré-categorização para os registros, isso significa que não é preciso um atributo como alvo principal FILHO (2007).

Ao se buscar resultados através de mineração de dados é comum a utilização de várias técnicas ou uma combinação entre elas de modo a obter os melhores resultados (COSTA et al., 2013). Neste trabalho, foram escolhidos os principais classificadores encontrados nos trabalhos de EDM: árvore de decisão, Naive Bayes, vizinho mais próximo - KNN, redes neurais, máquina de vetores de suporte (SVM) e a floresta aleatória - *Random Forest*.

2.1.3.1 Vizinho mais próximo - (KNN)

O KNN é um algoritmo que tem como objetivo realizar a classificação através da atribuição de rótulos que representam k amostras mais próximas através de votação. Para determinar a proximidade entre vizinhos é calculada a distância entre pontos dentro de um determinado espaço. Cada tupla representa um ponto em um espaço n-dimensional. Desta forma, todas as tuplas de formação são armazenadas num espaço padrão de n dimensões. Quando uma dada tupla é desconhecida, um classificador k-vizinho mais próximo procura o espaço padrão para as tuplas de treinamento k que estão mais próximas da tupla desconhecida. Estas tuplas de treinamento k são os k "vizinhos mais próximos" da tupla desconhecida (HAN; KAMBER; PEI, 2011). As fórmulas mais comuns para esta representação são distância Euclidiana e a distância de Manhattan.

2.1.3.2 Máquina de Vetores de Suporte *(Support Vector Machines)* - SVM

Os algoritmos de *Support Vector Machine* (SVM) ou Máquina de Vetores Suporte são algoritmos mais utilizados nos estudos de aprendizado de máquina e mineração de dados. Foi proposto pelo russo Vladmir Vapnik em 1979. A SVM é baseado na teoria da aprendizagem estatística e baseia-se no princípio da separação ótima de classes. Assim o objetivo dos algoritmos de SVM é determinar os limites de decisão que produzem uma separação ótima de classes por meio da minimização de erros (NASCIMENTO et al., 2009).

A SVM procura encontrar a linha hiperplana de separação entre duas classes. Este também busca maximizar a distância entre os pontos mais próximos. As SVMs têm a restrição de dados, que por sua vez, devem ser numéricos contínuos. Este modelo não é facilmente interpretável, e a seleção dos parâmetros adequados pode ser difícil (HÄMÄLÄINEN; VINNI, 2011).

2.1.3.3 Árvores de decisão - *Decision Tree*

Este algoritmo se tornou amplamente utilizado para inferências indutivas. O algoritmo ID3 (PADMAPRIYA; VELMURUGAN, 2014) é um dos mais populares por ser capaz de gerar uma árvore genérica. A árvore de decisão já foi utilizado para aplicar diagnósticos médicos e avaliar o risco de concessão de crédito financeiro .

O algoritmo trabalha semelhante a uma árvore e em cada nó é escolhido o atributo dos dados que mais efetivamente particiona o seu conjunto de amostras em subconjuntos de forma a tender a uma categoria ou a outra. O critério de particionamento é o ganho de informação normalizado (diferença em entropia). Isto implica dizer que o algoritmo consiste em um conjunto de condições, organizados em uma estrutura hierárquica. Este algoritmo é um modelo classificador de predição que, ao possuir condições satisfeitas desde a raiz da árvore até atingir suas folhas, que vai representar em um rótulo de classe. As árvores de decisão são consideradas modelos de fácil compreensão, porque um processo de raciocínio pode ser dado para cada conclusão, exceto se a árvore obtida é muito grande (uma série de nós e folhas) (ROMERO et al., 2008).

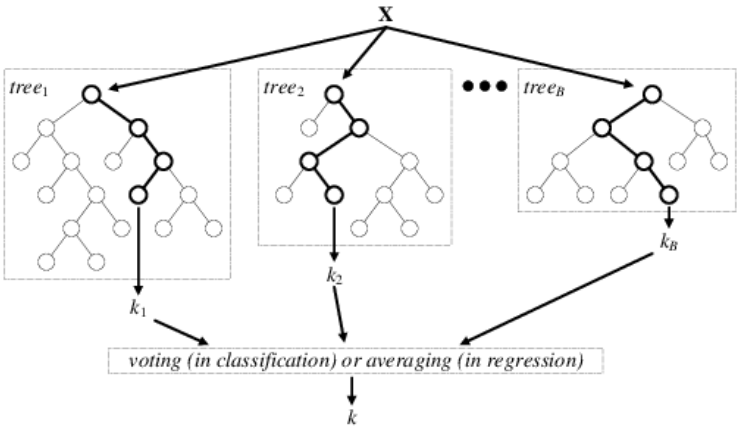
2.1.3.4 Floresta Aleatória - *Random Forest*

O *Random Forest* (RF) ou floresta aleatória é um algoritmo que também possui função classificatória, introduzido por Breiman em 2001. O RF é uma combinação de várias árvores de decisão para obter melhores predições. As RFs são obtidas através de *bootstrapping, aggregating ou bagging*, um método para gerar múltiplas versões de um preditor, inclusive estas versões são construídas a partir de reamostras do conjunto original, obtidas através de sorteio simples e com reposição (BASTOS; NASCIMENTO; LAURETTO, 2013).

É um algoritmo de mineração de dados que geralmente traz resultados eficazes. A

facilidade de sua abordagem o torna um dos algoritmos com maior utilização, porque é simples e pode ser usado para tarefas de classificação e regressão. Este procedimento extrai casos aleatoriamente a partir de conjuntos de dados de treinamento originais e os conjuntos são usados para construir cada uma das árvores de decisão que compõe a RF. Cada árvore classificadora é apontada como um componente preditor. A RF constrói sua decisão por meio da contagem dos votos dos componentes preditores em cada classe e, em seguida, seleciona a classe vencedora em termos de número de votos acumulados (CAMILO; SILVA, 2009).

Figura 2 – Arquitetura do modelo *Random Forest*.



Fonte: (VERIKAS et al., 2016)

2.1.3.5 Redes Neurais - *Neural Networks*

Neural Network (NN) ou Redes Neurais é um modelo matemático baseado na estrutura neural dos organismos inteligentes. Uma rede neural é composta por unidades de processamento que possuem canais de conexão associadas a pesos. Dessa forma a rede neural aprende com seu ambiente e melhora seu desempenho. Existem alguns modelos de Redes Neurais, entre eles a *Multilayer Perceptron*. Este é baseado em um modelo mais simples com uma camada chamado de Perceptron que permite a separação linear. A *Multilayer Perceptron* possui várias camadas que permite outros tipos de separação.

Uma rede neural artificial é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades, geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma rede neural artificial vem das interações entre as unidades de processamento da rede.

Para a utilização desse algoritmo é necessário um longo período de treinamento, além de ajustes de parâmetros. A interpretação deste modelo é de maior complexidade em comparação a outros algoritmos e não é possível identificar de maneira objetiva a relação das entradas e das

saídas. Entretanto as redes neurais, por sua vez, conseguem superar valores errados e podem identificar padrões para os quais nunca foram treinados (CAMILO; SILVA, 2009).

2.1.3.6 Naive Bayes

O Naive Bayes é um algoritmo probabilístico baseado no "teorema de Bayes". Este assume a redução da complexidade computacional para uma multiplicação de probabilidades, sendo um algoritmo simplificado. Uma de suas características é não considerar a correlação entre os atributos, o que o torna um algoritmo moderado. Este algoritmo possibilita as maiores acurácias na predição de performance acadêmico. (SAA, 2016).

Segundo Roy e Garg (2017a) o algoritmo Naive Bayes é de fácil implementação enquanto consegue manter um nível de acurácia alto. Para que este algoritmo tenha um bom nível de acurácia é necessário um repositório de dados de tamanho considerável, pois a sua precisão decresce consideravelmente em *datasets* pequenos. Os resultados podem ser ótimos de acordo com a base de dados.

2.1.4 *Education Data Mining*

Segundo Baker e Yacef (2009) a *Education Data Mining* (EDM) faz o uso da mineração de dados (DM), aprendizado de máquina e algoritmos estatísticos em um contexto de dados educacionais. Para Sukhija, Jindal e Aggarwal (2015), a EDM trabalha na direção de melhor procurar esses repositórios para desenvolver uma compreensão do processo subjacente e usuários. Isso é feito para fazer o uso dos dados combinados através de diversas técnicas de mineração de dados e algoritmos de associação para ajudar a otimizar a prática educacional em benefício do usuário final.

De acordo com Baker e Yacef (2009) a EDM não se limita somente a mineração de dados e realiza suas inferências em diferentes áreas, que vão desde o aprendizado de máquina a psicometria. A EDM tem impactado a pesquisa educacional nos últimos anos e sua abordagem abrange aplicações de várias áreas como: *e-learning*; auto-estudo; e sistemas de tutoria inteligente.

Do ponto de vista prático, a mineração de dados educacionais permite extrair conhecimento a partir dos dados dos estudantes. Esta informações podem ser usados para diversos objetivos, como: validar e avaliar um programa de sistema educacional, melhorar a qualidade dos processos educacionais e estabelecer a base para um aprendizado mais eficaz. Ideias semelhantes já foram aplicadas com sucesso, especialmente nos negócios, para aumentar os lucros de vendas (ALGARNI, 2016).

As campos referentes a EDM podem ser divididas em mineração e visualização estatística. A Tabela 1 mostra os principais cenários onde é possível aplicar a mineração de dados educacionais e os seus objetivos.

Tabela 1 – Cenários de mineração de dados educacionais

Categoria	Objetivos	Aplicações
Predição	Desenvolver um modelo para prever algumas variáveis em relação a outras. As variáveis preditoras podem ser constantes ou oriundas de um conjunto de dados.	Identificação de alunos em risco. Entender os resultados educacionais do aluno.
Clustering	Agrupar uma quantidade específica de dados com base nas características. O número de clusters pode ser diferente com base no modelo e os objetivos do processo de agrupamento.	Encontrar semelhanças e diferenças entre alunos ou escolas. Categorizar o comportamento do aluno.
Associação	Extrair o relacionamento entre duas ou mais variáveis no conjunto de dados.	Encontrar as relações de alunos que evadem. Descoberta de associações curriculares do curso. Descoberta de estratégias pedagógicas mais efetivas.
Descoberta com modelos	Destina-se a desenvolver um modelo de conhecimento, como um componente no modelo mais abrangente de previsão ou mineração de relacionamento.	Descoberta das relações entre os comportamentos dos alunos e o contexto das variáveis; Análise da questão de pesquisa em todo o contexto.
Destilação de dados para análise humana	O principal objetivo deste modelo é encontrar uma nova maneira de permitir que os pesquisadores identifiquem ou classifiquem os dados facilmente.	Identificação humana de padrões na aprendizagem do aluno ou comportamento; Rotular dados para uso em desenvolvimento posterior do modelo de previsão.

Fonte: (ALGARNI, 2016)

2.1.5 Técnicas de Avaliação de Modelos

2.1.5.1 Métricas de Avaliação para Algoritmos

Para que seja possível analisar o desempenho de um modelo treinado devemos selecionar métricas que nos possibilitem o cálculo efetivo da predição ou a correta separação das classes. Dentro destes cenários podemos destacar a estrutura denominada matriz de confusão (HAN; KAMBER; PEI, 2011). A matriz de confusão apresenta os resultados obtidos através de uma matriz bidimensional, que correlaciona linha e coluna para cada classe, de acordo a Tabela 2. Cada

elemento mostra o número de instâncias corretas ou incorretamente classificadas considerando-se o conjunto de testes utilizados. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Tabela 2 – Matriz de confusão

	Classe A1 prevista	Classe A2 prevista
Classe A1 prevista	Taxa de verdadeiro positivo (VP)	Taxa de falso negativo (FN)
Classe A2 prevista	Taxa de falso positivo (FP)	Taxa de verdadeiro negativo (VN)

A partir da matriz de confusão demonstrada na Tabela 2 podemos obter medidas para avaliar o desempenho do modelo EDM. Uma dessas medidas é a acurácia, medida esta que, mede a taxa de acerto global através do número de classificações corretas dividido pelo número total de instâncias dos dados a serem classificados conforme abaixo.

Figura 3 – Fórmula da acurácia.

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN}$$

A acurácia da classificação em um determinado conjunto N é medida pela taxa de classificação, que define a proporção de registros corretamente classificados no contexto deste mesmo conjunto (HÄMÄLÄINEN; VINNI, 2011) É importante destacar que além da acurácia existem outras medidas para avaliação de modelos EDM de acordo com (HAN; KAMBER; PEI, 2011). Estas medidas podem ser visualizadas na Tabela 3 a seguir:

Tabela 3 – Medidas para avaliação de modelos EDM

Fonte: (HAN; KAMBER; PEI, 2011)

Medida	Fórmula
Acurácia, taxa de reconhecimento	$\frac{VP+VN}{P+N}$
Taxa de erro, taxa de classificação incorreta	$\frac{FP+FN}{P+N}$
Sensibilidade, taxa de verdadeiro positivo, recobrimento	$\frac{VP}{P}$
Especificidade, taxa de verdadeiro negativo	$\frac{VN}{N}$
Precisão	$\frac{VP}{P+N}$
F, F1, F-score, média harmônica de precisão, recobrimento	$\frac{2 \times precisao}{precisao + recobrimento}$
Média Geométrica.	$MG = \sqrt{VP \times VN}$

2.1.5.2 Validação Cruzada

A validação cruzada é um procedimento para estimar o desempenho generalizado em um contexto. Este é o método mais comumente usado para avaliação de desempenho preditivo de um modelo, dado de antemão ou quando é desenvolvido por um procedimento de modelagem.

Os dados geralmente são divididos em duas partes com base nesta divisão, por um lado, o treinamento é feito enquanto o desempenho preditivo é testado na outra parte. O esquema de treinamento e teste funciona igualmente bem para modelos de classificação de aprendizado de máquina. Treinamos um modelo usando algumas instâncias do conjunto de dados e deixamos algumas instâncias fora dele para testar o modelo após ter sido treinado. Este é o princípio subjacente na validação cruzada.

Assim, a validação cruzada é amplamente aceita na comunidade de mineração de dados e aprendizado de máquina além de servir como um procedimento padrão para o propósito da seleção de modelo ou seleção de procedimento de modelagem (YADAV; SHUKLA, 2016).

2.1.5.3 *Receiver Operating Characteristic Graphs* - Gráficos ROC

É um método gráfico para avaliar, organizar e selecionar sistemas de diagnóstico e/ou predição (PRATI et al., 2008). A análise de ROC foi introduzida no aprendizado de máquina e mineração de dados como uma ferramenta útil e poderosa para a avaliação de modelos de classificação apresentando classes negativas e positivas, ou simplesmente, 0 e 1 (BRADLEY, 1997), (SPACKMAN, 1989). É particularmente útil em áreas onde existe uma grande desproporção entre as classes ou quando diferentes custos/benefícios devem ser levados em consideração para diferentes erros/sucessos de classificação. A análise ROC também tem sido utilizada para construir e refinar modelos.

2.1.6 Preparação dos dados

2.1.6.1 Seleção de Atributos para EDM

Um fator determinante para que os algoritmos voltados para EDM tenham índices de aproveitamento relevantes é a seleção de dados que facilitem a aprendizagem, uma vez que, os números dos atributos crescem de acordo com a dimensão. Este aumento que se apresenta de forma exponencial está diretamente ligada a primeira consequência da Maldição da Dimensionalidade (JEONG; KIM; CHOI, 2007).

A Maldição da Dimensionalidade pode ser descrito como todos os eventos que induzem dados com alta dimensão e que geralmente trazem efeitos desfavoráveis para o desempenho dos algoritmos de aprendizagem (VERLEYSEN; FRANCOIS, 2005). Uma das principais consequências deste fenômeno é a degradação das métricas que avaliam os modelos EDM.

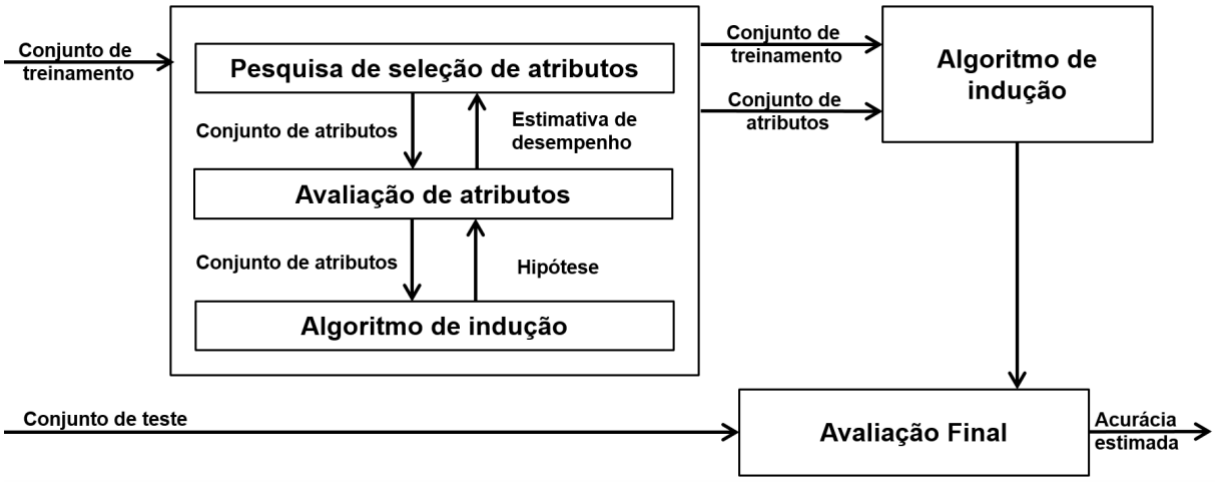
Para reduzir a ocorrência destes eventos podemos aplicar a redução da dimensionalidade.

Esta técnica tem como principal objetivo a redução de atributos para uma concepção reduzida dos dados com o mínimo de perda da informação (HAN; KAMBER; PEI, 2011). Portanto a redução da dimensionalidade dos dados através da retirada de elementos que não contribuem para o aprendizado refinam o algoritmo, potencializando assim, os resultados obtidos. É importante ressaltar que esta técnica tem por consequência uma representação mais simplificada e facilmente interpretável do conceito alvo, destacando os atributos que são mais relevantes (WITTEN; FRANK; HALL, 2011).

A problemática da seleção de subconjunto de atributos é justamente encontrar quais destes dados originais podem formar o subconjunto de forma a um algoritmo de indução, quando executado contendo apenas esses atributos, gere um classificador com a maior acurácia possível. A seleção de subconjunto de atributos é uma tarefa de mineração voltada a redução da dimensionalidade dos dados, onde são detectados e removidos atributos irrelevantes, fracamente relevantes ou redundantes (HAN; KAMBER; PEI, 2011). Para esta abordagem é possível adotar duas técnicas principais: *wrapper* e *filter*.

Para a técnica *wrapper* o algoritmo de seleção do subconjunto de dados existe como um revestimento sobre o algoritmo de indução (JOHN; KOHAVI; PFLEGER, 1994). O algoritmo de seleção do subconjunto de dados efetua pesquisas somente no contexto do subconjunto pré determinado utilizando o próprio algoritmo de indução como parte fundamental da avaliação do subconjuntos de dados.

Figura 4 – Demonstração de seleção de atributos através da abordagem *wrapper*.



Fonte: (KOHAVI; JOHN, 1997)(adaptado)

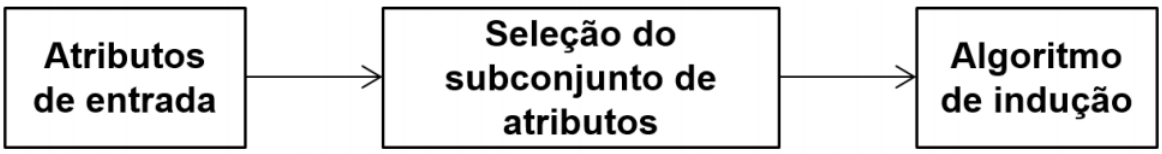
A abordagem *wrapper*, mostrada na Figura 4, usa o algoritmo de indução em dois conjuntos de dados distintos, sendo eles, o conjunto de treinamento e outro para avaliação do modelo. O subconjunto de atributos que possuir melhor desempenho é selecionado como conjunto no qual o algoritmo de indução realizará a varredura. Em seguida será utilizado a validação cruzada sob condição x para avaliar todo o contexto trabalhado. A técnica da validação

cruzada é uma metodologia experimental voltada a testes amplamente utilizada para avaliar algoritmos de classificação. Este procedimento divide o conjunto de atributos em K conjuntos menores e independentes entre si, todos com tamanhos semelhantes (WITTEN; FRANK; HALL, 2011).

Com esta definição é realizado K experimentos, e em cada experimento, o subconjunto de número K é removido. A partir deste ponto o algoritmo é treinado com os dados remanescentes, para em seguida, ser testado no subconjunto que foi selecionado. O resultado da validação cruzada garante que cada subconjunto teve seu teste independente o que assegura a imparcialidade e integridade da técnica (WITTEN; FRANK; HALL, 2011).

A técnica *filter* tende a selecionar os dados através de pré-processamento, desta forma, sua principal desvantagem é não levar em consideração as implicações geradas no subconjunto de atributos selecionados na avaliação do algoritmo de indução (KOHAVI; JOHN, 1997). Apesar desta característica, pode-se destacar que a abordagem *filter* tem por definição custo computacional menor quando comparado com a abordagem *wrapper*.

Figura 5 – Demonstração de seleção de atributos através da abordagem *filter*.



Fonte: (KOHAVI; JOHN, 1997)(adaptado)

2.1.6.2 Normalização

A normalização é uma técnica de escalonamento ou uma técnica de mapeamento ou um estágio de pré-processamento. É onde podemos encontrar uma nova gama a partir de uma gama existente. Pode ser muito útil para a previsão ou finalidade de previsão. Assim, para manter a grande variação de predição e previsão, a técnica da normalização é necessária para torná-los mais próximos. Existem algumas técnicas de normalização, dentre elas, a técnica que fornece a transformação linear na faixa original de dados que é chamada de normalização Min-Mix (PANDA; NAG; JANA, 2014).

2.1.6.3 Smote - *Synthetic Minority Oversampling Technique*

O desempenho dos algoritmos de aprendizado de máquina é normalmente avaliado usando precisão preditiva. No entanto, isso não é apropriado quando os dados estão desequilibrados e/ou os resultados variam acentuadamente. Um conjunto de dados é considerado desequilibrado se as classes não são representadas de maneira aproximadamente igual (CHAWLA et al., 2002).

A técnica de balanceamento SMOTE primeiro seleciona uma instância de classe minoritária e encontra seus k vizinhos de classe minoritária mais próximos. A instância sintética é então criada escolhendo um dos k vizinhos mais próximos b aleatoriamente e conectando a e b para formar um segmento linear no espaço. Desse modo as instâncias sintéticas são geradas como uma combinação convexa das duas instâncias escolhidas a e b (HE; MA, 2013).

2.1.7 Ferramenta de apoio para tratamento de dados

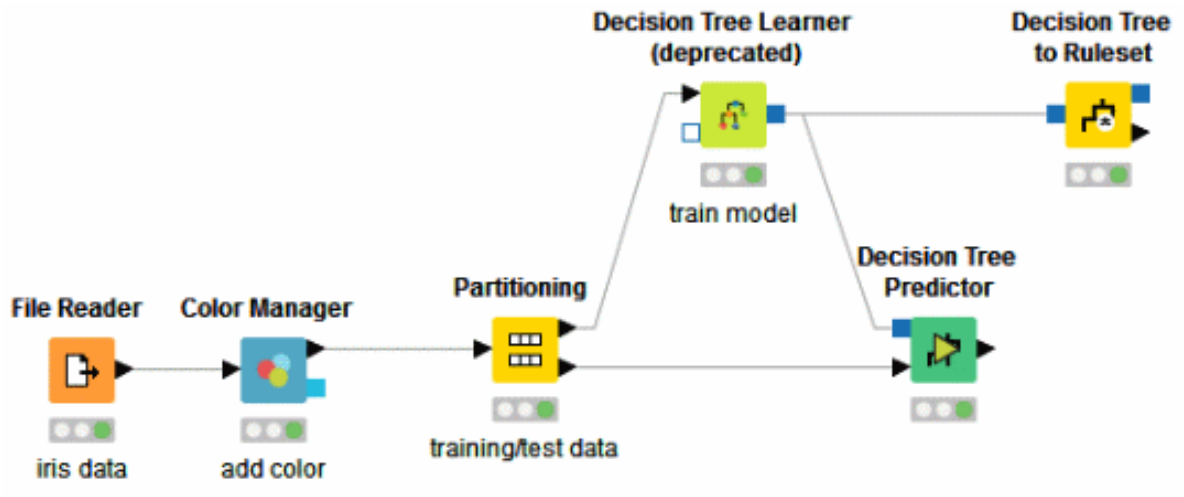
2.1.7.1 KNIME - *The Konstanz Information Miner*

O *Konstanz Information Miner* é um software modular baseado em nódulos construído a partir do ambiente de desenvolvimento integrado da IDE Eclipse. Cada nó executa o processamento dos dados sendo capaz de interagir com outros nós, permitindo a geração e a gravação de fluxos complexos de trabalho durante a manipulação da informação. O KNIME é uma plataforma poderosa de integração e análise preditiva. O software ajuda a documentar etapas realizadas no pré-processamento, análise estatística, modelagem estatística e análise preditiva. Outro ponto crucial é a sua colaboração de código aberto, onde os colaboradores são livres para desenvolver novas algoritmos, ferramentas e métodos de manipulação ou visualização de dados (BERTHOLD et al., 2009).

O desenvolvimento do KNIME começou em 2004 e foi liderado por uma equipe de engenheiros do Vale do Silício e da Universidade de Konstanz, na Alemanha, com Michael Berthold como líder do projeto. O software encontrou uso pesado em quimioinformática devido ao nível de automação necessário para explorar estruturas moleculares (BEISKEN et al., 2013). Atualmente, a ferramenta tem aproximadamente cinquenta por cento dos usuários provinientes de áreas bastante diferentes da ciência, medicina, negócios e ciências sociais, sendo por exemplo, utilizado em bancos privados em Zurique no gerenciamento de relacionamento com o cliente (WARR, 2012).

No KNIME, o usuário pode modelar fluxos de trabalho, que consistem em nós que processam dados, transportados por meio de conexões entre esses nós. Um fluxo geralmente começa com um nó que rastreia dados de alguma fonte de dados, que geralmente são arquivos de texto, mas os bancos de dados também podem ser consultados por nós especiais. Os dados importados são armazenados em um formato baseado em tabela interna que consiste em colunas com certos dados (extensíveis) tipo (inteiro, string, imagem, molécula, etc.) e um número arbitrário de linhas em conformidade com as especificações da coluna (BERTHOLD et al., 2009). Para esta pesquisa o KNIME foi utilizado principalmente para o pré-processamento das informações, análise estatística, modelagem e análise preditiva dos dados conforme a Figura 6.

Figura 6 – Fluxo de trabalho no KNIME



Fonte: (ANTUNES et al., 2018)

2.2 Revisão da Literatura

2.2.1 Mapeamento Sistemático

Este mapeamento sistemático contemplou os anos de 2010 até agosto de 2018 e é considerado a base fundamental que viabilizou a abordagem desta pesquisa. No momento de sua realização em agosto de 2018, o período considerado contempla intervalos mais recentes de modo a identificar quais as tendências atuais da área de mineração de dados quando aplicadas à evasão escolar. Nesta etapa, foram empregadas as técnicas e abordagens propostas por Kitchenham, et al. (2007) e Petersen, et al. (2008) no âmbito de procedimentos para a revisão e mapeamento sistemático. Pela especificidade do tema proposto, optou-se pela utilização de técnicas relacionadas a mapeamento sistemático, uma vez que, não se esperavam muitos estudos publicados nesta área de conhecimento.

2.2.1.1 Questões de Pesquisa

Este mapeamento sistemático orientou as pesquisas futuras, bem como a definição de quais algoritmos seriam melhor aplicados à evasão escolar. Para atingir este objetivo foram elaboradas quatro questões de pesquisa com o intuito de direcionar e categorizar os estudos deste trabalho.

QP1. *Quais são os países que mais publicam artigos sobre o tema?*
Com o intuito de direcionar os estudos no tema proposto, o questionamento acima tem por objetivo mapear os países que mais publicam artigos com este tema. Esta análise visa compreender se países emergentes estão desenvolvendo trabalhos que abordem a mineração de dados com o foco na evasão escolar.

QP2. *Quais modalidades de aprendizado aplicam a mineração de dados em temas de*

evasão escolar?
Esta questão tem por objetivo identificar quais modalidades de ensino aplicam algoritmos de mineração de dados para apoiar decisões voltados à evasão dos alunos. Esta informação refletirá se os problemas de evasão são mais frequentes em plataformas de aprendizado online, também conhecidos como EAD ou em modalidades presenciais.

QP3. *Quais são os propósitos para a utilização de mineração de dados para estudos voltados à evasão escolar?*
Este questionamento visa compreender o que os estudiosos buscam atingir quando utilizam a mineração de dados em suas pesquisas, visto que, estas técnicas se destinam a predição de eventos, análise comparativa entre desempenho de algoritmos, revisão de literatura, estudos de caso, análise de um conjunto de estudantes ou afins.

QP4. *Quais os principais algoritmos de mineração de dados utilizados para estudos relacionados à evasão escolar?*
Esta pergunta visa levantar quais são os algoritmos de mineração de dados mais utilizados durante a investigação ou desenvolvimento de pesquisas aplicadas à evasão escolar. Esta informação é crucial para alcançar os resultados esperados para cada investigação.

2.2.1.2 Fonte de Dados

Os estudos iniciais coletados que compuseram este mapeamento foram obtidos através da pesquisa em bases eletrônicas voltadas a publicações científicas que atenderam os seguintes critérios pré estabelecidos: (i) as bases dos trabalhos devem conter periódicos ou conferências na área de computação; (ii) as bases selecionadas devem permitir a pesquisa através de strings personalizadas ou palavras-chave que representem os termos abordados e (iii) as referidas bases devem permitir a leitura ou acesso completo aos documentos pesquisados. Dentre as fontes de pesquisa eletrônica disponíveis foram selecionadas:

- ACM Digital Library(http://dl.acm.org/);
- IEEEXplore(http://ieeexplore.ieee.org/);
- SpringerLink(https://link.springer.com/);
- Elsevier(https://www.elsevier.com/pt-br); e,
- WileyInterScience(http://www.sciencedirect.com)

2.2.1.3 Critérios para Inclusão/Exclusão

Com a finalidade de refinar os resultados iniciais da busca, foram definidos alguns critérios de inclusão e exclusão. Os critérios seguem detalhados na Tabela 4.

Tabela 4 – Modelo para inclusão e exclusão de estudos.

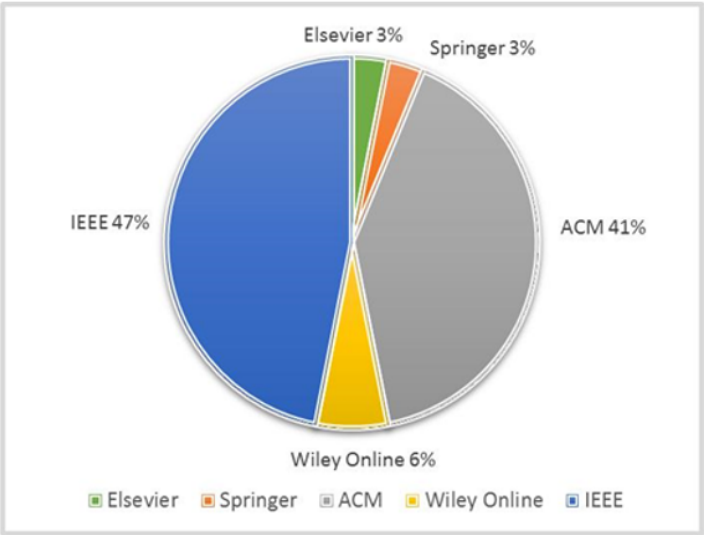
Critério	Descrição
Inclusão	Os artigos devem estar disponíveis na web. Os artigos devem envolver alguma técnica de mineração de dados. Data de publicação superior ao ano de 2010.
Exclusão	Artigos escritos em outras línguas além do inglês. Conjunto de slides, posters, demonstrações e afins. Estudos com finalidade ambígua. Estudos em que não sejam abordadas técnicas de mineração ou afins.

Após a definição da estratégia, a busca primária retornou 1583 artigos. Em seguida foi realizada a leitura dos títulos e resumos destes, o que reduziu o número inicial para 51 estudos selecionados. Em última aplicação dos critérios de inclusão e exclusão pré-estabelecidos, foram definidos 32 artigos para compor este mapeamento sistemático.

2.2.1.4 Resultados

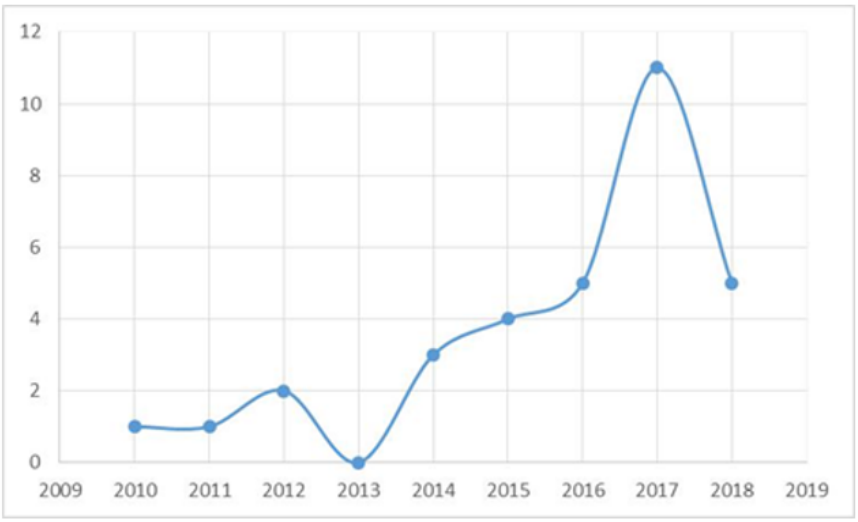
Para uma melhor visualização foram instituídas visões gerais das coletas e assim responder as perguntas já elencadas em 2.2.1.1. Após a realização do protocolo de mapeamento sistemático foram selecionados 32 trabalhos conforme a Figura 7.

Figura 7 – Contribuições de cada base de pesquisa.



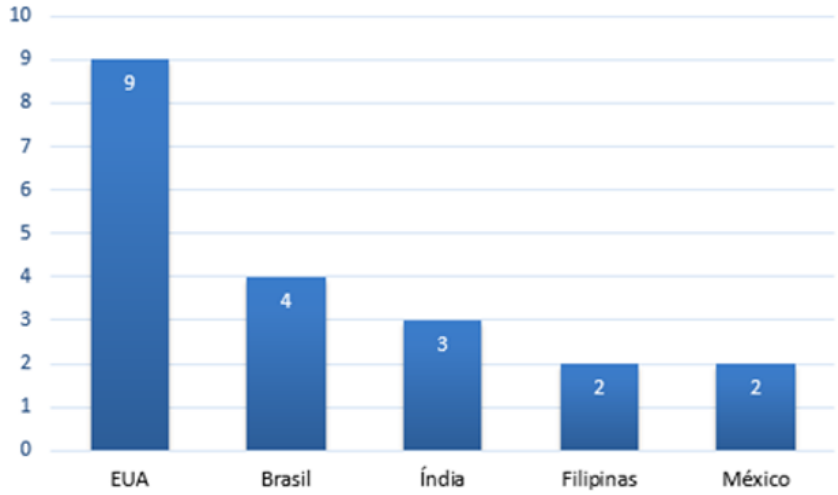
Em ordem decrescente temos o IEEE com 15 artigos (47%), em seguida temos o ACM com 13 publicações (41%), Wiley Online com 2 pesquisas (6%) e por último temos o Springer (3%) e Elsevier (3%) com 1 colaboração cada. Além desta contribuição percebeu-se que, durante o período analisado (2010 a 2018), houve um grande acréscimo nas publicações referente a área de mineração de dados aplicadas à evasão escolar, conforme ilustra a Figura 8. Onde é possível observar que o ano de 2017 foi o mais contributivo no âmbito de pesquisas para o tema proposto.

Figura 8 – Quantidade de publicações por ano.



QP1. *Quais são os países que mais publicam artigos sobre o tema?*
É possível visualizar que o país que mais se destaca em publicações de artigos na área é os Estados Unidos. Conforme podemos observar na Figura 9.

Figura 9 – Países que mais publicam sobre o tema.



Um dos objetivos de longo prazo de qualquer universidade nos EUA é a redução do atrito estudantil, pois cerca de um quarto dos estudantes abandona a faculdade após o primeiro ano de estudo. Este número aumenta para 50% até o final do quarto semestre letivo (AMERI et al., 2016). Em seguida é possível perceber que o Brasil vem demonstrando interesse em utilizar a mineração de dados para evitar a evasão de alunos. Este problema universitário é uma preocupação diária, no entanto, a identificação precoce dos alunos que tem um perfil evasivo pode ser uma solução. Desta forma é possível utilizar medidas corretivas para o discente antes que o mesmo considere o abandono do curso superior (MARTINS et al., 2017). Posteriormente, percebe-se que a Índia segue em destaque para publicações nesta área, pois uma educação melhor

desempenha um papel proeminente e muitas atividades são realizadas pelo governo e instituições de ensino para aumentar o valor de educação no país. A educação contribui com a elevação do status social, da saúde, do progresso econômico, para o sucesso da nação, conscientização sobre problemas ambientais e questões relacionadas (ATHANI et al., 2017). Em seguida identificamos que Filipinas é o quarto país a publicar artigos nessa área de pesquisa e que muitas instituições perseguem este objetivo (BENABLO et al., 2018). No México os estudos já são aplicados em instituições do ensino médio (MARQUEZ-VERA; MORALES; SOTO, 2013). Outros países como Austrália, Bósnia, China, Chipre, Espanha, Finlândia, Grécia, Indonésia, Japão, Lituânia, Reino Unido e Taiwan também contribuíram com 1 pesquisa respectivamente para a composição deste mapeamento sistemático.

QP2. *Quais modalidades de aprendizado aplicam a mineração de dados em temas de evasão escolar?*

A principal modalidade de ensino que aplicou mineração de dados para temas de evasão escolar foi o modelo presencial com 25 citações (69%). Já a modalidade EAD apresentou 11 estudos (31%) onde foram utilizadas algumas técnicas de mineração de dados. É importante ressaltar que destes 32 trabalhos que compuseram este mapeamento sistemático, 4 estudos apresentaram simultaneamente técnicas de mineração de dados aplicadas para cenários presenciais e de aprendizado à distância conforme pode-se observar na Tabela 5.

Tabela 5 – Modalidade de Ensino e EDM.

Modalidades de Ensino	Citações
Online / EAD / MOOC	(ROMERO; VENTURA, 2017),(AL-SHABANDAR et al., 2017), (ANGRA; AHUJA, 2017),(DWIVEDI; ROSHNI, 2017), (LIANG; LI; ZHENG, 2016),(SUKHIJA; JINDAL; AGGARWAL, 2015), (BOGARÍN et al., 2014),(CALDERS; PECHENIZKIY, 2012), (IHANTOLA et al., 2015),(RODRIGUES; ISOTANI; ZÁRATE, 2018), (KOSTOPOULOS; KOTSIANTIS; PINTELAS, 2015)
Presencial	(MARQUEZ-VERA; MORALES; SOTO, 2013), (BALANIUK et al., 2011), (LIU; CHEN, 2010), (ANGRA; AHUJA, 2017), (ATHANI et al., 2017), (BRAJKOVIĆ; RAKIĆ; KRALJEVIĆ, 2018), (DEEPAK et al., 2016), (GOPALAKRISHNAN et al., 2017), (KOSTOPOULOS et al., 2017), (MARTINS et al., 2017), (PRISTYANTO; SETIAWAN; ARDIYANTO, 2017), (ROY; GARG, 2017a), (SLIM et al., 2014), (SUKHIJA; JINDAL; AGGARWAL, 2015), (PUARUNGROJ et al., 2018), (AMERI et al., 2016), (CALDERS; PECHENIZKIY, 2012), (IHANTOLA et al., 2015), (LAKKARAJU et al., 2015), (LAURÍA et al., 2012), (NURHUDA; ROSITA, 2017), (RAGAB et al., 2014), (RUSTIA et al., 2018),(MANHÃES; CRUZ; ZIMBRÃO, 2014), (BENABLO et al., 2018)

Estudos feitos em universidades demonstram que as desistências de alunos universitários podem ser atribuídas a uma série de razões como, por exemplo, scores em exames pré-universitários, o apoio financeiro familiar e nível de educação dos pais. Todos estes fatores contribuem para a persistência ou não do discente (GOPALAKRISHNAN et al., 2017). A falta de motivação e mudanças de vida inesperadas podem atrasar a formatura dos alunos e alavancar a desistência dos estudantes. É importante ressaltar que discentes que não se formam podem sobrecarregar os recursos alocados as universidadese demais unidades escolares (LAKKARAJU et al., 2015).

Massive Open Online Courses (MOOCs) tornaram-se uma plataforma educacional alternativa que permitem aos alunos de localizações geográficas distintas a mesma qualidade de aprendizado através da web. *Coursera*, *HarvardX* e *KhanAcademy* são alguns exemplos de MOOCs. Desde 2012 esta modalidade é utilizada de maneira intensiva pelas maiores Universidades. Estudos apontam que o ensino a distância atraiu muitos participantes aos cursos oferecidos devido a remoção das barreiras financeiras, geográficas e educacionais (AL-SHABANDAR et al., 2017). Apesar dos cursos MOOCs serem considerados uma revolução no campo educacional, uma das principais preocupações é a alta taxa de evasão. Isto significa que a grande maioria dos alunos que se inscrevem no começo dos cursos não chega a etapa final. A evasão pode ser detectada a partir de diferentes contextos como, por exemplo, a visualização de vídeos, postagens em fóruns entre outros (LIANG; LI; ZHENG, 2016).

QP3. *Quais são os propósitos para a utilização de mineração de dados para estudos voltados à evasão escolar?*

Dentre os 32 estudos selecionados neste mapeamento sistemático, 17 destes (53,12%) apontaram que a EDM é mais citada na literatura em estudos que avaliam a acurácia de algoritmos previamente estipulados além de expor estes resultados em relação a dados de alunos matriculados em unidades de ensino (MANHÃES; CRUZ; ZIMBRÃO, 2014). Em seguida, 7 destes trabalhos (21,87%) se propuseram a analisar informações existentes através de algoritmos para identificação perfis de alunos evasivos ou a fomentar a construção de modelos preditivos de forma macro, isto é, definir se haverá evasão ou não do discente avaliado (SLIM et al., 2014). Dos trabalhos levantados 6 artigos (18,75%) apresentaram revisões de literatura sobre o tema e em último podemos citar os estudos de caso que obtiveram 3 citações (9,37%) ao longo do desenvolvimento deste estudo. Podemos considerar que apesar dos algoritmos de mineração de dados já serem academicamente uma opção para evitar a evasão escolar ainda temos poucas aplicações em cenário real.

A tomada de decisão é muito importante para a educação, seja para melhoria da qualidade de ensino ou melhor utilização de recursos disponíveis. Na Tabela 6 é possível visualizar os resultados obtidos.

Tabela 6 – Propósito de uso para EDM.

Objetivos Gerais	Citações
Predição de eventos	(ANGRA; AHUJA, 2017), (BRAJKOVIĆ; RAKIĆ; KRALJEVIĆ, 2018), (DWIVEDI; ROSHNI, 2017),(SLIM et al., 2014), (PUARUNGROJ et al., 2018),(NURHUDA; ROSITA, 2017), (BENABLO et al., 2018)
Análise comparativa	(LIU; CHEN, 2010), (AL-SHABANDAR et al., 2017), (ATHANI et al., 2017),(DEEPAK et al., 2016), (GOPALAKRISHNAN et al., 2017), (KOSTOPOULOS et al., 2017), (LIANG; LI; ZHENG, 2016),(MARTINS et al., 2017), (PRISTYANTO; SETIAWAN; ARDIYANTO, 2017), (AMERI et al., 2016), (BOGARÍN et al., 2014), (KOSTOPOULOS; KOTSIANTIS; PINTELAS, 2015), (LAKKARAJU et al., 2015), (LAURÍA et al., 2012), (RAGAB et al., 2014),(RUSTIA et al., 2018), (MANHÃES; CRUZ; ZIMBRÃO, 2014),
Estudo de caso	(MARQUEZ-VERA; MORALES; SOTO, 2013), (BALANIUK et al., 2011), (IHANTOLA et al., 2015)
Revisão de literatura	(ROMERO; VENTURA, 2017), (ROY; GARG, 2017a), (SUKHIJA; JINDAL; AGGARWAL, 2015), (IHANTOLA et al., 2015), (RODRIGUES; ISOTANI; ZÁRATE, 2018) (CALDERS; PECHENIZKIY, 2012)

QP4. *Quais os principais algoritmos de mineração de dados utilizados para estudos relacionados à evasão escolar?*

Em ordem decrescente observa-se o SVM com 12 citações (37,50%), Naive Bayes com 9 citações (28,12%), o C4.5 e a Regressão Logística com 6 citações (18,75%), em seguida a Floresta Aleatória com 5 (15,62%) e por último ambos Redes Neurais (NN) e *Multilayer Perceptron* com 4 citações (12,50%) respectivamente cada um.

Outros algoritmos também foram citados em menor número e portanto não foram considerados para agregação à Tabela 7, entre eles podemos citar o *Linear Regression*, *CHAID*, *GBM*, *Adtree*, *Genetic Algorithm*, *Bagging*, *Multi-Task Feature Learning* – MTFL, *Heuristic Miner*, *Reptreee Adboost*.

A escolha do algoritmo é muito importante para qualquer trabalho na área de mineração de dados (ATHANI et al., 2017). A aplicabilidade dos algoritmos pode se mostrar versátil para a tratativa dos problemas e é comum que vários destes sejam aplicados em um mesmo trabalho para lapidar corretamente as informações que se deseja extrair (MANHÃES; CRUZ; ZIMBRÃO, 2014). A avaliação dos resultados elencou os 7 algoritmos que recorrentemente foram citados durante a análise dos 32 trabalhos que compuseram esta pesquisa conforme demonstra a Tabela 7.

Tabela 7 – Algoritmos recorrentes em temas de evasão.

Algoritmo	Citações
C4.5	(AL-SHABANDAR et al., 2017), (PUARUNGROJ et al., 2018), (KOSTOPOULOS; KOTSIANTIS; PINTELAS, 2015), (LAURÍA et al., 2012), (RAGAB et al., 2014), (RUSTIA et al., 2018)
Regressão Logística	(AL-SHABANDAR et al., 2017), (LIANG; LI; ZHENG, 2016), (MARTINS et al., 2017), (BOGARÍN et al., 2014), (LAURÍA et al., 2012), (RUSTIA et al., 2018)
SVM	(MARQUEZ-VERA; MORALES; SOTO, 2013), (ATHANI et al., 2017), (BRAJKOVIĆ; RAKIĆ; KRALJEVIĆ, 2018), (DEEPAK et al., 2016), (GOPALAKRISHNAN et al., 2017) (AL-SHABANDAR et al., 2017), (LIANG; LI; ZHENG, 2016), (PRISTYANTO; SETIAWAN; ARDIYANTO, 2017), (LAURÍA et al., 2012), (RUSTIA et al., 2018), (MANHÃES; CRUZ; ZIMBRÃO, 2014), (BENABLO et al., 2018)
Multilayer Perceptron	(BALANIUK et al., 2011), (KOSTOPOULOS et al., 2017), (RAGAB et al., 2014), (MANHÃES; CRUZ; ZIMBRÃO, 2014)
Naive Bayes	(MARQUEZ-VERA; MORALES; SOTO, 2013), (BRAJKOVIĆ; RAKIĆ; KRALJEVIĆ, 2018), (DEEPAK et al., 2016), (KOSTOPOULOS et al., 2017), (MANHÃES; CRUZ; ZIMBRÃO, 2014), (RAGAB et al., 2014), (RUSTIA et al., 2018), (PRISTYANTO; SETIAWAN; ARDIYANTO, 2017) (AL-SHABANDAR et al., 2017),
Redes Neurais	(AL-SHABANDAR et al., 2017), (ATHANI et al., 2017), (NURHUDA; ROSITA, 2017), (RUSTIA et al., 2018)
Floresta Aleatória	(AL-SHABANDAR et al., 2017), (LIANG; LI; ZHENG, 2016), (MARTINS et al., 2017), (MARTINS et al., 2017), (RAGAB et al., 2014)

2.2.1.5 Trabalhos Relacionados

Existem muitos trabalhos publicados na área de *Education Data Mining* visto que a possibilidade de prever o abandono dos alunos dos cursos de ensino superior tornou-se uma pesquisa importante e desafiadora para as universidades (KOSTOPOULOS; KOTSIANTIS; PINTELAS, 2015). No entanto, apesar das contribuições destas pesquisas ainda existem lacunas que podem ser preenchidas através de abordagens pouco exploradas. Deste modo, os trabalhos relacionados aqui elencados, são exemplos de pesquisas que obtiveram destaque no meio acadêmico e que de certo modo são referências quando se busca aprofundamento no tema de pesquisa proposto.

Shaun et al. (2011) apresentam os principais métodos que descrevem a mineração de dados educacionais como predição, agrupamento, mineração e descobertas com modelos além de apontar o grande potencial destas técnicas para otimizar a qualidade de ensino no Brasil. Através do uso de métodos da EDM sugere melhorar os modelos de conhecimento do estudante em vários diferentes domínios como ensino de língua estrangeira, geometria, química, física e muitos outros. No entanto, seu trabalho não aplica as técnicas informadas em uma abordagem.

Manhaes et al. (2012) já utilizam técnicas de mineração para identificar alunos que não conseguem concluir o curso pretendido no curso superior na Universidade Federal do Rio de Janeiro. Para alcançar este objetivo foi realizado 3 experimentos com o intuito de determinar qual algoritmo apresenta melhor acurácia no cenário em que foi aplicado. Os algoritmos selecionados são *OneR* (OR), *JRip* (JR), *Decision Table* (DT), *Simple Cart* (SC), *J48* (J48), *Random Forest* (RF), *Simple Logistic* (SL), *Multilayer Perceptron* (MP), *Naive Bayes* (NB), *Bayes Net* (BN). A amostra contemplou 7304 alunos de graduação e foram utilizados seis algoritmos de classificação com acurácia acima de 80% nos resultados obtidos. Deste trabalho concluí-se, que a acurácia dos classificadores e a taxa de erro são fortemente influenciadas pelos vieses da base de dados e permite que a universidade não utilize apenas dados estatísticos na análise do problema da evasão. No entanto, este trabalho não apresenta como os dados foram tratados ou como identificar o momento em que o aluno desistirá do curso.

Kantorski et al. (2016) realizaram um trabalho através da mineração de dados educacionais para visualizar e possibilitar perspectivas de modo a mitigar a evasão de alunos. Neste contexto foi realizado dois estudos experimentais aplicada a dois cursos de graduação presencial da Instituição: Curso de Zootecnia e Curso de Administração. Os experimentos foram realizados através da ferramenta WEKA com cinco algoritmos distintos (*J48*, *IBk*, *CART*, *Naive Bayes*, *Multilayer Perceptron*), mas somente três (*J48*, *CART* e *Naive Bayes*) apresentaram resultados mensuráveis. Como resultado foi possível gerar uma lista de prováveis alunos que deixariam de realizar a matrícula no ano posterior ao cursado. Este experimento alcançou 98% de acurácia para a previsão de desistência e mais de 70% para alunos que realmente evadiram do curso. No entanto o método proposto extrai informações pessoais, acadêmicas, sociais e econômicas. Em cenários onde não existe sistemas integrados é inviável ter acesso a todas a estas informações dos discentes.

Júnior et al. (2015) ressaltam que a evasão escolar no ensino superior é um fenômeno em crescimento e que recentemente se tornou um foco de preocupação para estudiosos de diferentes áreas. A pesquisa explorada propôs uma abordagem computacional para auxílio à tomada de decisão através de uma seleção de melhores atributos para a tarefa de classificação, considerando classes que "haverá evasão" e "não haverá evasão". Dentro desse contexto foram empregados os classificadores de árvore de decisão (*J48*), baseado em regras (*JRip*), máquina de vetores de suporte (*SVM*, com a implementação *SMO*), redes neurais artificiais (*Multilayer Perceptron*), métodos de conjunto de classificadores (*Random Forest*) e o classificador K vizinhos mais próximos (*IBk*). Foi constatado que 80% das evasões concentram-se até o 3º período do curso, independente do total de períodos do curso (6, 8 ou 10 períodos). Os resultados experimentais contribuíram com inferências para apoio à tomada de decisão de gestores educacionais em níveis estratégicos, tático e operacional. No entanto, além de não ser possível identificar qual aluno evadirá do curso selecionado não existe uma forma facilitada para compreender estes dados, como por exemplo, uma lista de possíveis alunos evasivos. Este cenário dificulta qualquer ação que evite a desistência do aluno do curso.

Marquez-Vera, Morales e Soto (2013) propuseram a aplicação de técnicas de mineração de dados para predizer o abandono escolar, contemplando 670 estudantes de ensino médio em Zacatecas, México. Foram empregados 10 algoritmos para esta atividade dentro da ferramenta WEKA. Dentre eles são utilizados 5 algoritmos de indução: *JRip*, que é um aprendiz de regras propostos; *NNge*, que é um algoritmo semelhante ao vizinho mais próximo; *OneR*, que usa o atributo de erro mínimo para predição de classe; *Prism*, que é um algoritmo para induzir regras modulares; e *Ridor*, que é uma implementação da regra *Ripple-Down Learner*. Os outros 5 algoritmos são compostos de árvore de decisão: *J48* , que é um algoritmo para gerar árvore de decisão, *C4.5* ajustada ou não ajustada; *Simple Cart*, que implementa poda de custo-complexidade; *ADTree*, que é uma árvore de decisão alternativa; *Random Tree*, que considera K atributos escolhidos aleatoriamente em cada nó da árvore e *REPTree*, que aprende rapidamente a árvore de decisão. Os resultados obtiveram acurácias dentre 75% a 98% dentre os dez classificadores selecionados para este estudo de caso quando aplicados em diferentes contextos no mesmo experimento. Para o desenvolvimento e análise deste trabalho, as notas dos alunos foram utilizadas como ponto chave em relação a outros atributos presentes. Como conclusão os autores declaram que algoritmos de classificação podem ser utilizados para predição de desempenho escolar. No entanto, apesar destas contribuições ainda não é possível identificar claramente quando o aluno evadirá e tampouco identificar em que classe/matéria poderá ocorrer a evasão.

No cenário da Universidade de São Paulo, a razão entre alunos que finalizaram o curso e o total de vagas oferecidas em 2013 foi de apenas 12,8%, com uma média de alunos concluintes por docente ativo de 1,24. Na Escola de Artes, Ciências e Humanidades, a taxa estimada de evasão em 2013 atingiu 12%, enquanto no Bacharelado em Sistemas de Informação esse número foi de 13,3%. De modo que podemos afirmar que, em média, a cada 180 alunos que ingressam no curso durante o ano, aproximadamente 24 estudantes não conseguem se graduar (DIGIAMPIETRI; NAKANO; LAURETTO, 2016). Neste cenário um estudo de caso foi aplicado considerando apenas as cinco disciplinas específicas ministradas por professores do curso no primeiro ano do Bacharelado em Sistemas de Informação, a saber: ACH0021-Tratamento e Análise de Dados/Informações; ACH2001-Introdução à Programação; ACH2011-Cálculo I; ACH2002-Introdução à Análise de Algoritmos; e ACH2012-Cálculo II. Neste momento foi aplicado o algoritmo classificador *Rotation Forest* utilizando validação cruzada em dez subconjuntos, considerando tanto disciplinas individuais como subconjuntos das disciplinas analisadas. A acurácia deste estudo de caso obteve aproveitamento de 65% a 92% de acerto. Como principal contribuição desta pesquisa é proposto a análise de um conjunto de matérias correspondente a um semestre letivo. No entanto não é citado quais atributos dos alunos foram utilizados para o modelo proposto.

Adeodato et al. (2004) propuseram um modelo com base em uma comparação completa de desempenho entre regressão logística e redes neurais (*Multilayer Perceptron*), realizada em uma amostra de 180.000 exemplos por meio de um processo de validação cruzada estratificado

de 30 vezes com intervalos de confiança de 9,5% nas medidas de desempenho, com o objetivo de buscar alunos com potencial risco de evasão já no final do segundo semestre do curso, ou seja, durante o segundo período letivo. Para este estudo foram selecionados e analisados seis cursos da Universidade Federal de Pernambuco. Dentre todas as informações disponíveis, apenas vinte e uma variáveis foram derivadas dos dados originais. Esta seleção teve como objetivo identificar atributos comuns a todos os cursos. Entre os atributos selecionados estão: notas médias referentes ao 1º e 2º semestres cursados; a variação referente a taxa de reprovação do 1º para o 2º período; a taxa de aprovação obtidas nos dois semestres do curso dentre outros. O estudo obteve o desempenho de 84% de aproveitamento ao utilizar a área sob curva ROC com escala de 0 a 1. Apesar de todas as contribuições desta pesquisa ainda não é possível identificar a evasão do aluno através da disciplina cursada.

Lam-On e Boongoen (2014) salientam que os bancos de dados costumam adicionar muitos atributos redundantes que, consequentemente, contribuem diretamente a condições que podem degradar o desempenho de alguns algoritmos de classificação. Para mitigar a baixa acurácia em decorrência de variáveis repetidas, os autores propuseram a realização de *Extract, Transform, Load* - ETL antes de aplicar os algoritmos de mineração de dados selecionados. Após a aplicação de seis técnicas de transformação de dados e quatro algoritmos de classificação foi possível obter a acurácia de aproximadamente 92% durante a pesquisa. Neste contexto a pesquisa focou-se em limpeza dos dados e não se concentrou em predizer o momento em que o aluno evadirá do curso selecionado.

Antunes (2010) combina diferentes técnicas de mineração como a classificação e regras de associação para prever o momento de falha dos alunos. Os resultados experimentais foram aplicados em um conjunto de dados de alunos da graduação do Instituto Superior Técnico de Lisboa, matriculados na disciplina de Fundamentos de Programação. Para este estudo foram utilizados os algoritmos *J48*, *CART* e *ASAP* dentro da ferramenta WEKA. No entanto neste trabalho não é possível identificar quais são os alunos que falharão no curso.

Roy e Garg (2017b) utiliza os algoritmos *Naive Bayes*, *J48*, *decision tree* e *MLP* para classificar os alunos que evadiram do curso. Nesta pesquisa são utilizados mais de 30 atributos dos alunos. A acurácia destes algoritmos obteve entre 51% a 73% de aproveitamento. É importante ressaltar que neste estudo foram utilizados dados médicos pertencentes aos estudantes para compor o modelo proposto. No entanto, nesta pesquisa a acurácia dos algoritmos não é tão assertiva além de ser necessário vários atributos para compor o modelo computacional.

Musso et al. (2013) em sua pesquisa utilizou o algoritmo *ANN* em uma amostra total que incluiu 864 estudantes universitários, de ambos os sexos (masculino 45,4%; feminino 54,6%), com idades entre 18 e 25, recém-matriculados no primeiro ano em várias disciplinas diferentes (psicologia, engenharia, medicina, direito, comunicação social, negócios e marketing), em três universidades privadas na Argentina, durante os anos acadêmicos de 2009-2011. Ao todo, 67,8% da amostra tinha 17 a 20 anos, 24,7% tinha 21-25 anos e 7,5% tinha mais de 25 anos. Os

alunos da amostra provinham de escolas particulares religiosas (48,5%), escolas particulares não religiosas (19%), escolas particulares bilíngues (15,4%), escolas públicas de ensino médio (15%) e 2,1% de escolas comunitárias internacionais. Todos os dados dos alunos (preditores) foram coletados no início do ano acadêmico correspondente, e a variável dependente (GPA) foi coletada no final do mesmo ano letivo. Os resultados alcançados neste estudo permitiram identificar a influência específica de cada conjunto de variáveis de input nos diferentes níveis de desempenho escolar (alto e baixo rendimento), por um lado, e processos comuns a todos os alunos. Por outro, uma contribuição importante dessa abordagem preditiva é a constatação de que as mesmas variáveis têm efeitos diferentes em cada grupo de alunos, definindo padrões específicos para cada nível de desempenho. Embora a contribuição de cada variável em um padrão particular carregue um peso preditivo relativamente pequeno, é o efeito combinado do padrão de variáveis que explica um modelo de desempenho acadêmico mais baixo ou mais alto. No entanto esta abordagem não é viável em cenários com muitos discentes, justamente por conta dos diferentes efeitos observados em grupos com baixo e alto rendimento.

Osmanbegovic e Suljic (2012) utilizam *Naive Bayes*, *MLP*, *J48* em um conjunto de 257 estudantes da universidade da Tuzla, do curso de Economia com o apoio da ferramenta WEKA. A acurácia alcançadas por estes algoritmos ficou entre 72% a 75% de aproveitamento. Nesta pesquisa é destacado o uso do *Naive Bayes* como algoritmo principal para a construção de modelos preditivos, no entanto a abordagem utilizada ainda é muito inicial quando comparados com outros estudos já citados anteriormente.

Para Vasconcelos (2019) a evasão é, certamente, um dos grandes problemas que afligem as instituições de ensino em geral, uma vez que as perdas ocasionadas pelo abandono do aluno são desperdícios sociais, acadêmicos e econômicos. Neste trabalho foi realizado duas análises experimentais dos algoritmos de mineração de dados mais utilizados na área de educação, com intuito de avaliar o que melhor se adequa ao contexto de abandono do ensino em duas instituições federais. Foram planejados e executados dois experimentos controlados "in vivo", para comparar a eficácia dos classificadores selecionados. Em seguida, foi realizado um estudo de caso com interface criada para aplicar o algoritmo que obteve a melhor eficácia. Os resultados evidenciaram diferenças significativas entre os algoritmos utilizados, apesar do SVN possuir a maior média das métricas de eficácia, os algoritmos MLP e *Random Forest*, respectivamente, obtiveram resultados semelhantes de acurácia (85,38%, 84,40% e 84,13%). Apesar da alta acurácia dos experimentos este estudo apresenta experimentos controlados "in vivo" que é uma abordagem diferente da proposta desta pesquisa que é exploratória.

Medina et al. (2020) comparam os algoritmos rede bayesiana e árvore de decisão em um cenário de mineração de dados educacional (EDM). Os dados foram coletados junto a 500 alunos de graduação de uma universidade particular de Lima, Peru. Os resultados indicam que as redes bayesianas têm um desempenho melhor do que as árvores de decisão baseadas em métricas de precisão, exatidão, especificidade e taxa de erro. Neste estudo, a precisão das redes

bayesianas chega a 67,10% enquanto para árvores de decisão é de 61,92%. A abordagem utiliza como principais atributos informações sociais do aluno, como por exemplo, renda familiar, escolaridade dos pais, entre outros. No entanto este estudo não consegue atingir uma acurácia expressiva.

Hegde e Prageeth (2018) apresentam uma metodologia para prever o abandono escolar usando o algoritmo de classificação *Naive Bayes* em linguagem R. Este estudo aproveita alguns fatores para compor os seus atributos como, por exemplo, fatores acadêmicos, fatores demográficos, fatores psicológicos, problemas de saúde, opinião do professor e comportamento do aluno. Dos 54 atributos selecionados variás técnicas EDM foram aplicadas, sendo a principal destas, a redução da dimensionalidade. Nesta abordagem o autor aborda tópicos através de perguntas orientadas. Nesta pesquisa o algoritmo obteve acurácia de 72%. Apesar desta abordagem apresentar acurácia relevante, a quantidade de atributos, bem como o cunho destes como, por exemplo, dados médicos sugere integrações de sistemas onde seja possível coletar estas informações pessoais.

Perez, Castellanos e Correal (2018) apresentam um estudo de caso focado na detecção de evasão de alunos de graduação em Engenharia de Sistemas após 7 anos de matrícula em uma universidade colombiana. Os algoritmos árvores de decisão, regressão logística e *Naive Bayes* foram comparados para propor a melhor opção. Além disso, o *Watson Analytics* também foi utilizado para estabelecer a usabilidade do serviço para pessoas sem formação em tecnologia. Os 43 atributos analisados são oriundos de 802 alunos matriculados no Programa de Ciência da Computação em uma universidade privada em Bogotá, Colômbia. Dentre estes atributos também estão presentes dados sobre condições financeiras do aluno. O algoritmo árvores de decisão apresentou o maior valor de acurácia atigindo 94% de acerto, em seguida o algoritmo regressão logística apresentou 92% de acurácia e o *Naive Bayes* apresentou 87% de acerto neste estudo. Apesar da alta acurácia dos algoritmos, alguns atributos utilizados nesta pesquisa são de dificil acesso principalmente a entidades públicas quando não possuem sistemas integrados.

2.2.1.6 Considerações Finais

Business Intelligence, *Data Mining* e *Education Data Mining* podem ser utilizados de maneira conjunta em ambientes educacionais. Ainda neste contexto é possível utilizar ferramentas e softwares, como o Pentaho e o Knime, para facilitar a interpretação e visualização destas informações a cada nova interação/manipulação dos dados.

Além das ferramentas e técnicas computacionais foi necessário compreender como o *Education Data Mining* é retratado em outros países, como a modalidade de ensino, seja EAD ou presencial, pode influenciar o estudo EDM, quais algoritmos são mais utilizados em ambientes acadêmicos entre outros. Para tal, foi realizado um mapeamento sistemático para direcionar/estabelecer a melhor forma de avançar nesta pesquisa e determinar quais são os algoritmos mais utilizados para esta abordagem.

A partir do mapeamento sistemático foi possível compreender como o *Education Data Mining* é abordado em trabalhos já publicados na área, quais foram as principais contribuições e quais são os principais limitantes de cada trabalho em destaque. Muitas pesquisas na área estão focadas em definir qual algoritmo obtém melhor desempenho através da acurácia da técnica selecionada (MARTINS et al., 2017). No entanto, esta abordagem pode apresentar resultados variáveis de acordo com os dados/atributos existentes nas instituições analisadas. Outros estudos visam identificar alunos que efetivamente abandonarão o curso selecionado (MACHADO et al., 2015). Contudo, esta abordagem é ineficaz quando se deseja identificar quando o aluno apresentará evidências de evasão e não auxilia os professores no apoio aos alunos com esses sintomas.

Desta forma, o objetivo deste trabalho é apresentar uma abordagem que visa uma análise aprofundada do contexto evolutivo do aluno por disciplina por meio de técnicas EDM. Esta abordagem visa reduzir a evasão escolar e aumentar a retenção de alunos nas instituições, além de auxiliar gestores e professores a identificar antecipadamente alunos com possibilidade de evasão das disciplinas e, no futuro, do curso selecionado. Para demonstrar o uso desta abordagem, um estudo de caso também é realizado.

Além da abordagem computacional, este estudo considera a situação individual de cada instituição, o histórico das disciplinas, o departamento responsável pela matéria a ser cursada, o histórico do aluno e a grade obrigatória que o discente deve realizar a cada semestre ou período. Estas informações são medidas por meio de atributos que facilitam a análise. Em geral, os resultados apresentados ou os novos atributos criados podem ser interpretados sob diferentes pontos de vista.

3

Cenário de Aplicação do Estudo de Caso

Esta Seção apresenta a modelagem dos dados adotada para esta pesquisa, o processo de ETL realizado para a transformação e atualização dos dados, e os principais atributos disponíveis para a pesquisa exploratória. É retratado como estas informações contribuíram para a análise preliminar e em seguida para a abordagem que avalia o desempenho de alunos em disciplinas.

3.1 Pentaho - BI

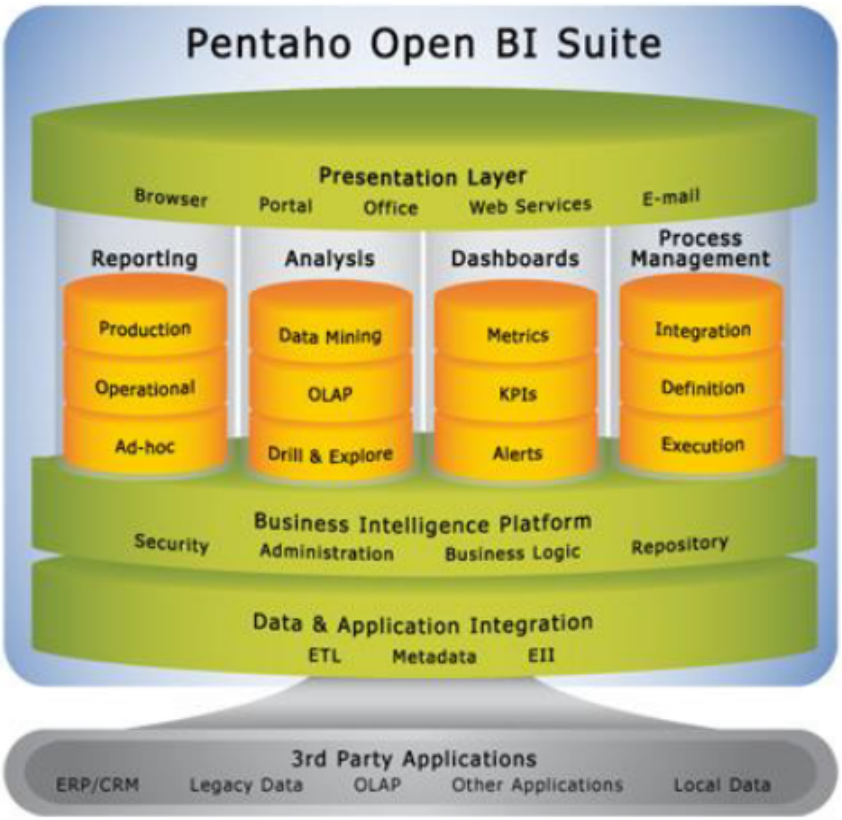
Em trabalho anterior foi realizado um experimento onde foi disponibilizado a ferramenta Pentaho para apoio à tomada de decisão junto a coordenação de cursos do departamento de computação (NETO, 2016). Este estudo disponibilizado em 2016 realizou a análise de dados experimentais até o ano de 2015 com o intuito de gerir *reports* a partir de informações já consolidadas.

O Pentaho é a versão *open source* que oferece no mesmo pacote tudo o que é necessário para desenvolvimento de uma solução BI. É também uma das mais conhecidas soluções entre os usuários de softwares livres. A Pentaho Community oferece suporte a relatórios, ferramentas de análise, integração de dados para ETC, OLAP, *data mining*, *dashboards* simples, suporte a *big data* e ferramentas de gerenciamento simples (NETO, 2016).

O Pentaho *Community* também oferece um ambiente com autenticação, regras para usuários e *WEB services*. O servidor integrado gerencia relatórios, integrações entre serviços e *workflow* do processo. Cada usuário possui perfil próprio, sendo possível a personalização de dashboards e outros relatórios. Suporta vários repositórios como Microsoft SQL Server, Oracle, PostgreSQL, MySQL, Ingres, HSQL e vários outros.

A Figura 10 mostra um diagrama da arquitetura do Pentaho. É possível visualizar de maneira geral como funciona o processo de entrada, processamento e saída. É possível notar a grande quantidade de recursos incorporados à solução (NETO, 2016).

Figura 10 – Arquitetura do Pentaho

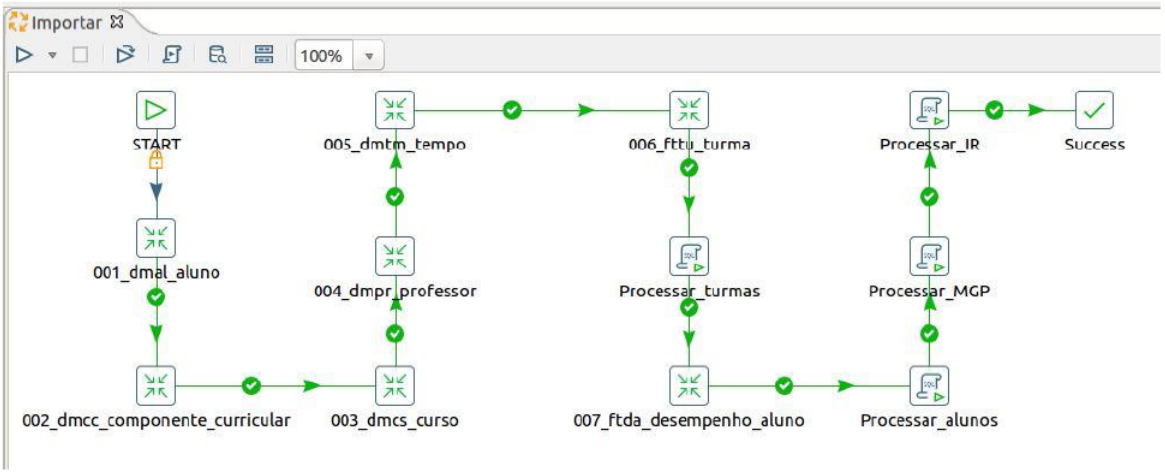


Fonte: (NETO, 2016)

3.2 Processo de ETL

Para realizar o processo de carga, é necessário utilizar o *Pentaho Data Integrator* (PDI) para a execução das transformações. Para melhor compreensão do processo de ETL, foi criado um *job* que permite visualizar todo o processo de transformação. A Figura 11 mostra o *job* que executa as transformações e alimenta o SGBD do Pentaho.

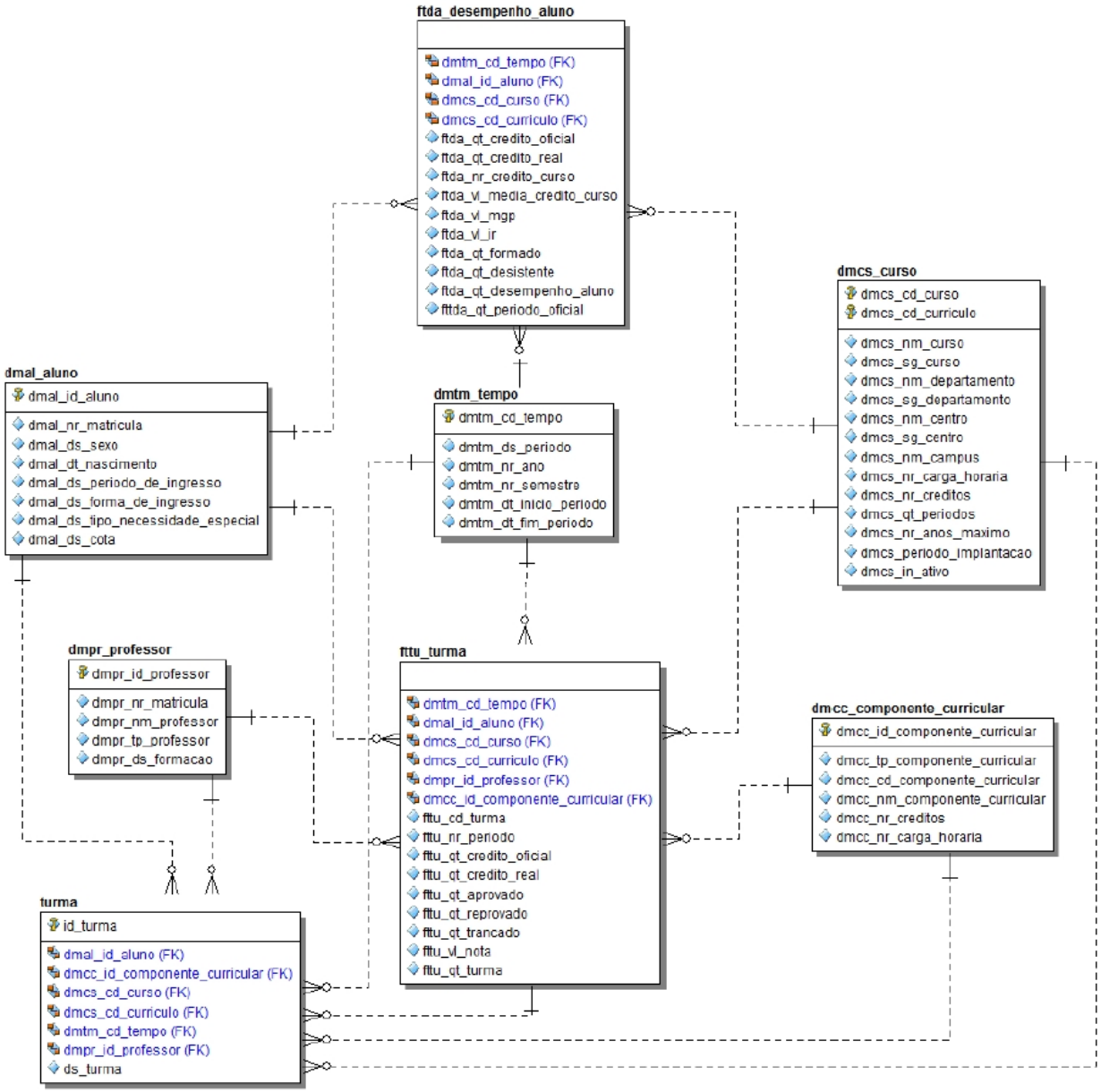
Figura 11 – *Job* do *Pentaho Data Integrator*.



3.3 Organização dos dados

Neste momento é utilizado o esquema estrela, onde os fatos são cercados por relações com as dimensões do problema. Este termo é usado pois a estrutura resultante parece uma estrela. Normalmente existe uma tabela central, a tabela de fatos, e um conjunto de tabelas menores, as tabelas de dimensões, dispostas em um padrão radial em torno da tabela de fatos. Esta estrutura do SGBD da Figura 12 é alimentada através do *job* do Pentaho apresentado na Figura 11. Para garantir a privacidade dos alunos os dados são anonimizados e não há informação que identifique qualquer aluno. Estes dados foram criptografados através de *hash* MD5.

Figura 12 – Estrutura estrela adotada nesta pesquisa.



A tabela ftda_desempenho_aluno é a tabela que concentra as informações referentes aos discentes da instituição e a tabela fttu_turma reflete todas as informações referentes as matérias ofertadas. Através destas tabelas é possível extrair os atributos primários necessários para a criação/transformação de novos atributos. A tabela fttu_turma é dententora dos atributos fttu_aprovado e fttu_reprovado e possibilita identificar se o aluno obteve aprovação na disciplina observada. Neste contexto, fttu_turma se refere diretamente a cada disciplina ofertada pelo Dcomp. Algumas informações somente são obtidas a partir de um cruzamento entre as tabelas fttu_turma e ftda_desempenho_aluno como, por exemplo, o total de disciplinas em que o aluno foi aprovado/reprovado.

Neste estudo será proposto três atributos que representarão consecutivamente as grandezas "dificuldade média da turma", "dificuldade média do aluno" e "dificuldade média do período a ser cursada pelo aluno". Para que seja possível a criação destes, ambas as tabelas fttu_turma e ftda_desempenho_aluno serão amplamente manipuladas de modo a extrair as informações necessárias citadas em 5.3.1 a 5.3.3 que serão utilizadas na Seção 5.3.4 para a transformação de dados e a criação dos novos atributos.

3.4 Quantidade de Alunos por ano

Após a Seção 3.1 e 3.2 o SGBD foi atualizado com dados mais recentes seguindo o processo apresentado nas Figuras 10 e 11, estes possibilitaram experimentos utilizando EDM para a posterior aplicação da abordagem que avalia o desempenho de alunos em disciplinas. Neste contexto, foram avaliados no total 4.017 alunos do Dcomp desde o ano de 2007 até 2018 e 49.013 registros em turmas disponíveis neste período. Abaixo podemos visualizar a quantidade de estudantes ingressos por ano:

Tabela 8 – Quantidade de ingressos Dcomp - UFS por ano.

Ano de Ingresso	Alunos
2007	100
2008	104
2009	207
2010	238
2011	246
2012	232
2013	338
2014	530
2015	656
2016	477
2017	430
2018	459

3.5 Seleção de Atributos

De acordo com a Seção 2.1.6.1 é necessário selecionar os atributos para EDM de modo a reduzir a dimensionalidade dos dados analisados. Deste modo este estudo utiliza a abordagem *filter*. Para a análise preliminar e para a abordagem que avalia o desempenho de alunos em disciplinas, somente alguns dos atributos apresentados no SGBD Pentaho serão utilizados. Para a análise preliminar serão utilizados os atributos: idade, gênero, forma de ingresso, disciplina de estudo e situação final da disciplina. Já para a abordagem que avalia o desempenho de alunos em disciplinas, utilizaremos os novos atributos DMT, DMA que serão detalhados na Seção 5, o atributo correspondente a aprovação do aluno, o atributo correspondente ao crédito real do aluno, os valores atributos DMT, DMA e crédito real do aluno com a função *autoBinner* do Knime.

3.6 Atributos Preexistentes

A lista de atributos a seguir foi obtida através do sistema de base de dados do Pentaho, que já são utilizadas pela Universidade Federal de Sergipe. Estes atributos contribuiram nos estudos apresentados seja na etapa preliminar, para a composição de novos atributos, para indução de atributos que contribuem para a análise da evasão e também para avaliar a eficácia do modelo adotado. A numeração dos atributos indicados abaixo segue a codificação utilizada no Tabela 9.

Atributo nº 1: Curso - Indica qual foi o curso selecionado pelo aluno, dentre eles: Ciência da Computação, Sistema de Informação ou Engenharia da Computação.

Atributo nº 2: Gênero - Indica o gênero do aluno ingressante, masculino ou feminino, de acordo com o seu registro civil.

Atributo nº 03: Tipo de cota - Indica em qual política de cotas o aluno foi selecionado no processo seletivo. A partir de 2017.2 são utilizados os seguintes tipos de cota:

• Grupo C - Não cotista.

• Grupos A, A1 - Oriundo de família com renda igual ou inferior a 1,5 salários-mínimos per capita e se enquadra no grupo PPI - Pretos, Pardos e Indígenas. Sendo A1 candidatos que possuem alguma deficiência.

• Grupos B, B1 - Oriundo de família com renda igual ou inferior a 1,5 salários-mínimos per capita e não se enquadra no grupo PPI. Sendo B1 candidatos que possuem alguma deficiência.

• Grupos D, D1 - Formado por candidatos cotistas, independentemente da renda e que se enquadram no grupo PPI - Pretos, Pardos e Indígenas. Sendo D1 candidatos que possuem alguma deficiência.

• Grupos E, E1 - Formado por candidatos cotistas, que não optam pela cota racial. Sendo E1 candidatos que possuem alguma deficiência.

Atributo nº 04: Coeficiente de MGP - A média ponderada de n valores com seus respec-

tivos n pesos é definida como a soma dos produtos dos n valores pelos seus respectivos pesos, dividindo-se este resultado pela soma dos pesos.

Atributo nº 05: Coeficiente IR - O Índice de Regularidade (IR) corresponde ao quociente entre a média dos créditos cursados pelo aluno a partir do seu ingresso na UFS no curso atual (CMA) e a média dos créditos que devem ser cursados para integralizar o currículo do curso no tempo padrão (CMC).

Atributo nº 06: Crédito Oficial - Crédito obtido após o aluno cursar a disciplina.

Atributo nº 07: Crédito Real - Crédito realmente cursado pelo aluno durante a vida acadêmica.

Atributo nº 08: Aprovado - Status do aluno após cursar a disciplina.

Atributo nº 09: Reprovado - Status do aluno após cursar a disciplina.

Atributo nº 10: Trancado - Status do aluno após cursar a disciplina.

Atributo nº 11: Média - Média do aluno obtida por disciplina.

Atributo nº 12: Formado - Status que identifica se o aluno conseguiu concluir a graduação escolhida.

Atributo nº 13: Turmas - Histórico das matrículas, aprovações, reprovações e trancamentos realizados em cada turma.

Tabela 9 – Atributos estudados durante os experimentos

Nº	Atributo	Tipo	Atributo Criado
01	Curso	Númerico	
02	Gênero	Categórico	
03	Tipo de cota	Categórico	
04	Coeficiente de MGP	Númerico	Sim
05	Coeficiente IR	Númerico	Sim
06	Crédito Oficial	Númerico	
07	Crédito Real	Númerico	
08	Aprovado	Númerico	
09	Reprovado	Númerico	
10	Trancado	Númerico	
11	Média	Númerico	
12	Formado	Númerico	
13	Turmas	Categórico	

4

Análise Preliminar da Evasão

Esta etapa tem como principal objetivo realizar, através de algoritmos de mineração de dados educacionais, uma análise experimental inicial que visa entender as razões que levam os estudantes a evadirem das unidades de ensino. Aqui é descrito um estudo preliminar da evasão, e reflete principalmente os resultados publicados na conferência internacional - ICALT 2019 (*The 19th IEEE International Conference on Advanced Learning Technologies*) sob o título *Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout* de O. Santos et al. (2019).

4.1 Metodologia

A metodologia utilizada para o trabalho em questão consiste, em termos de classificação, de uma pesquisa exploratória e descritiva. Dentro da pesquisa exploratória é destacado o mapeamento sistemático realizado, no qual, foram selecionados os classificadores: árvore de decisão, vizinho mais próximo (KNN), redes neurais, máquina de vetores de suporte (SVM), *Naive Bayes* e floresta aleatória (*Random Forest*). Segundo (MACHADO et al., 2015), esses algoritmos estão entre os mais utilizados em mineração de dados educacionais.

Neste momento, os algoritmos selecionados serão observados de acordo com a acurácia alcançada por cada um destes em relação aos atributos acadêmicos disponíveis no sistema SGBD. Esta etapa tem como objetivo principal compreender qual destes algoritmos possui maior desempenho, neste contexto, através de uma abordagem preliminar. É importante ressaltar que nesta abordagem inicial o período analisado (entre os anos 2010 a 2018) é diferente do período observado na Seção 5 (entre os anos 2007 a 2018), uma vez que, as informações presentes entre os anos de 2010 a 2018 são mais completos e com menos atributos faltantes. A partir deste ponto, será possível avançar a pesquisa para o estudo de caso, já citado na Seção 1, onde é realizado a interpretação dos resultados encontrados com as questões de pesquisa propostas.

Assim sendo, para atingir o objetivo da pesquisa inicial é necessário consolidar os dados já disponibilizados pelo *Pentaho* para análise experimental dos cursos de Ciência da Computação, Sistemas de Informação e Engenharia da Computação desde o 1º ao 6º semestre, respectivamente, dos anos de 2010 a 2018 da Universidade Federal de Sergipe. O período selecionado se destaca principalmente pela utilização do novo sistema SIGAA que reuniu dados de aplicações legadas sob uma única visão.

Para obter os resultados esperados, as etapas de processamento foram as seguintes:

1. Aquisição de dados da plataforma SIGAA da Universidade Federal de Sergipe (I.; ROSA, 2013);
2. ETL realizada pelo *Pentaho* e Pré-processamento de dados;
3. Aplicação dos seguintes algoritmos: Árvores de Decisão, Vizinho mais próximo - *KNN*, Redes Neurais, Máquina de Vetores de Suporte - *SVM*, *Naive Bayes* e Florestas Aleatórias - *Random Forest*;
4. Seleção de algoritmos com base na acurácia apresentada;
5. Avaliação dos resultados e sua análise.

Para todas as etapas citadas acima, a linguagem Python foi usada junto com o framework Anaconda e suas dependências (ANALYTICS, 2016).

Sobre a aquisição de dados: foram selecionados 23.690 alunos históricos dos programas de graduação do Departamento de Ciência da Computação da Universidade Federal de Sergipe. Entre eles, os cursos de Ciência da Computação (CC) obtiveram 12.079 registros, seguidos de Sistemas de Informação (SI) com 5.592 alunos e Engenharia da Computação (EC) com 5.389 alunos. Esses dados incluíram alunos e ex-alunos da instituição ao longo dos anos de 2010 a 2018, mas incluíram apenas registros do primeiro ao sexto semestre.

No pré-processamento dos dados, foi realizada uma análise semântica para correção de palavras incompletas, ambíguas ou duvidosas; etapas ETL também foram efetuadas com o objetivo de tratar as informações a partir dos arquivos CSVs e o SGBD foi normalizado para assegurar o comportamento e o aprendizado dos algoritmos selecionados. Os dados também foram revisados para garantir que não existissem elementos ausentes (A. FACELI K., 2011).

Sobre a etapa de seleção, foram selecionados alunos e ex-alunos (idade, gênero, forma de ingresso, disciplina de estudo e situação final da disciplina). Os demais atributos como qual turma, semestre, número de créditos e descrições foram descartados por não contribuírem para a precisão dos algoritmos (A. TOLOSI L.; T., 2010). Para facilitar a interpretação das informações, todas as strings foram convertidas em valores numéricos e os alunos foram agrupados em 3 faixas etárias: de 16 a 25 anos, de 25 a 35 anos e acima de 35 anos.

Para a etapa de classificação, as tarefas de aprendizagem supervisionada descritas anteriormente foram aplicadas aos dados pré-processados. Como um dos objetivos era avaliar o comportamento de diferentes técnicas a fim de obter o melhor desempenho, foi aplicada a validação cruzada com uma divisão de cinco subconjuntos e os parâmetros escolhidos para cada técnica podem ser vistos na Tabela 10.

Tabela 10 – Parâmetros selecionados para cada algoritmo.

Algoritmo	Parâmetros selecionados
KNN	Números de Vizinhos = 3
Decision Tree	Critério = Entropia; Número mínimo de divisões = 2; Amostras mínimas em um nó folha = 1
Naive Bayes	*Alpha* = 1; Aprenda as probabilidades anteriores = *True*
Neural Network	Número de camadas ocultas = 2; Tamanho da camada = 100; Ativação = ReLu Otimizador = Adam; Taxa de Aprendizagem = 0.001
Support Vector Machines	Função kernel RBF com Gamma = 0.2
Random Forests	Número de estimadores = 100; Profundidade Máxima = 2; Critério = Gini; Amostras mínimas em um nó folha = 1

Por fim, a discussão e análise de comparação entre os resultados foram feitas como última etapa.

4.2 Resultados Preliminares

Com relação ao melhor algoritmo referente a acurácia para todos os seis períodos do curso de CC foram Árvore de Decisão e *Random Forest*, ambos com 66% respectivamente. Já para o curso de SI a acurácia dos algoritmos *Random Forest*, *SVM* e Árvore de Decisão atingiu o marco de 70%. Para o curso de EC o melhor algoritmo foi a Árvore de Decisão com 72% de acurácia.

Tabela 11 – Acurácia média para os melhores algoritmos

Curso	Melhores Algoritmos	Acurácia Média
Ciência da Computação	- Árvore de Decisão - *Random Forest*	66%
Sistemas de Informação	- *Random Forest* - *SVM* - Árvore de Decisão	70%
Engenharia da Computação	- Árvore de Decisão	72%

O algoritmo que apresentou a acurácia mais baixa para os cursos de CC e EC foi o *KNN* com 63%. Já no contexto do curso de SI o *Naive Bayes* apresentou o resultado 65% que o deixa em última posição quando relacionado aos demais. Podemos observar todos os resultados consolidados através das Figuras 13 a 18 e as Tabelas 12 a 14.

Figura 13 – Acurácia no período 4 de SI

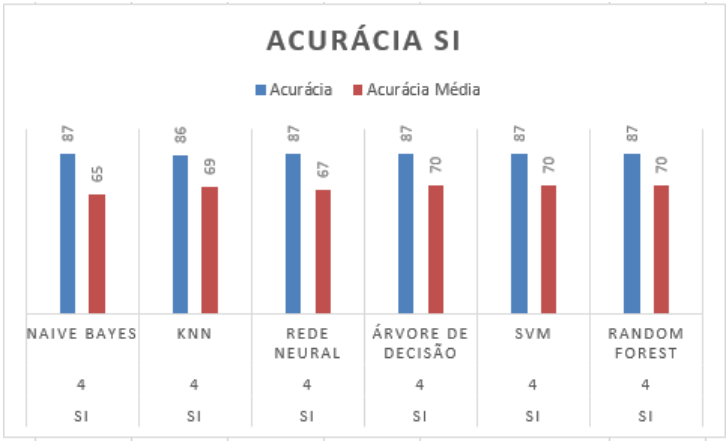


Figura 14 – Acurácia no período 6 de CC

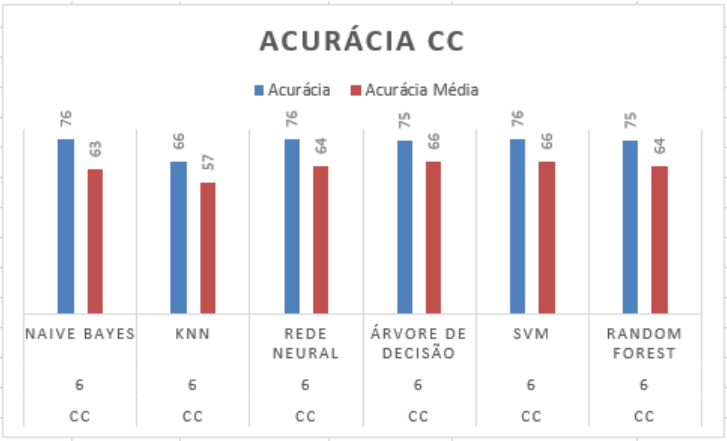


Figura 15 – Acurácia no período 4 de EC

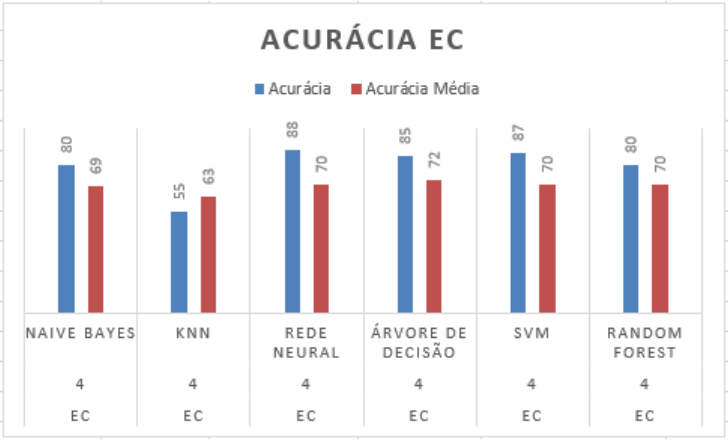


Figura 16 – Comparação rendimento SI por período

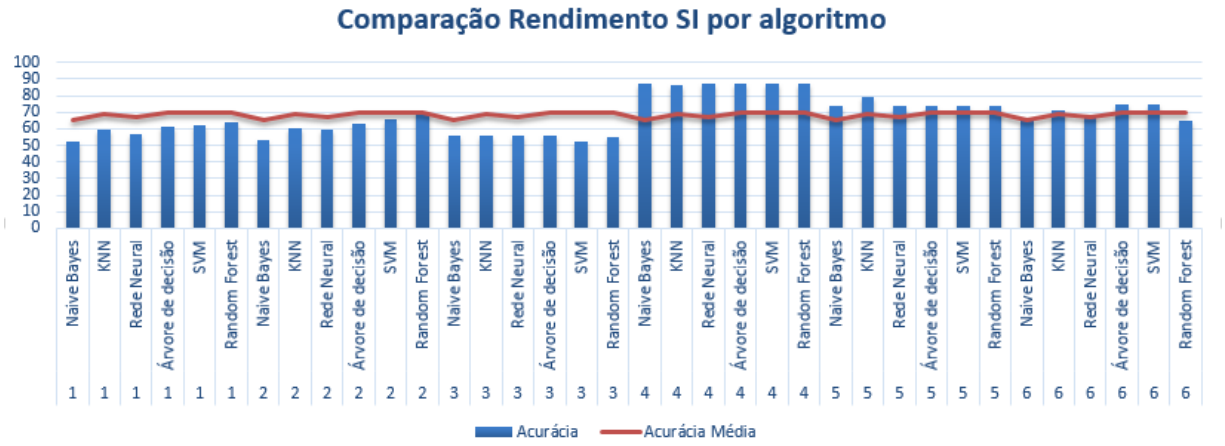


Figura 17 – Comparação rendimento EC por período

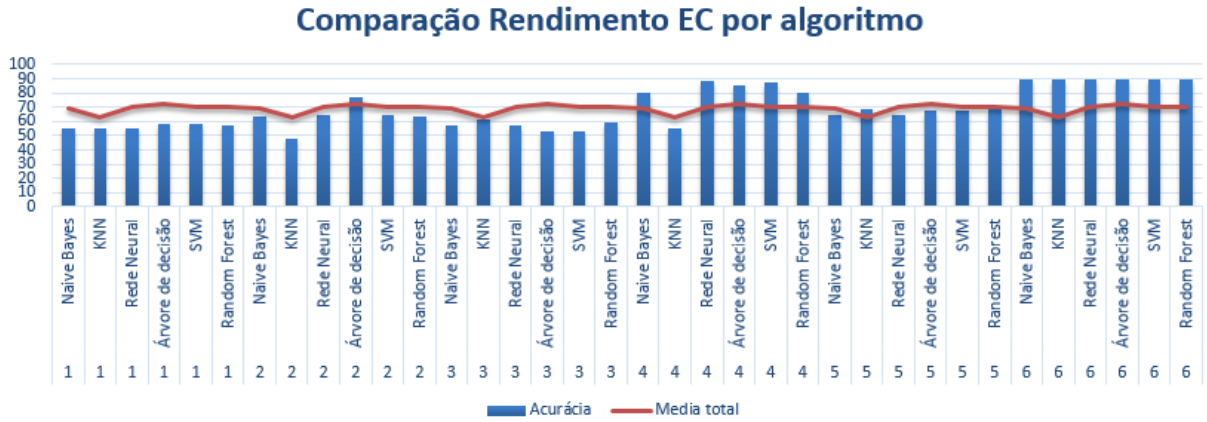


Figura 18 – Comparação rendimento CC por período

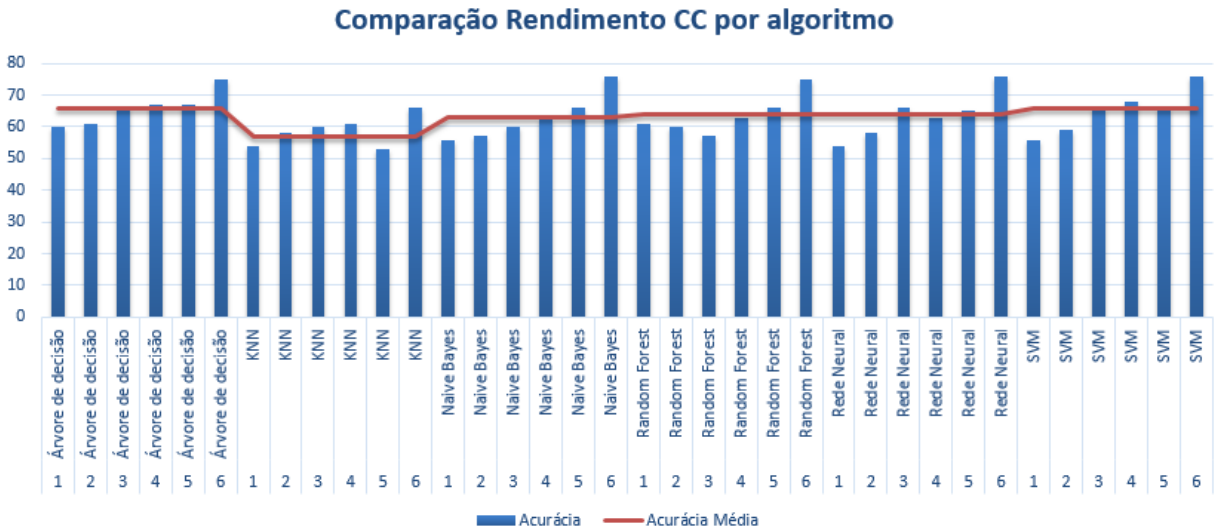


Tabela 12 – Análise de algoritmo por período SI

Curso	Período	Algoritmo	Acurácia	Acurácia Média
SI	1	Árvore de decisão	0.61	0.70
SI	2	Árvore de decisão	0.63	0.70
SI	3	Árvore de decisão	0.56	0.70
SI	4	Árvore de decisão	0.87	0.70
SI	5	Árvore de decisão	0.74	0.70
SI	6	Árvore de decisão	0.75	0.70
SI	1	KNN	0.59	0.69
SI	2	KNN	0.60	0.69
SI	3	KNN	0.56	0.69
SI	4	KNN	0.86	0.69
SI	5	KNN	0.79	0.69
SI	6	KNN	0.71	0.69
SI	1	Naive Bayes	0.52	0.65
SI	2	Naive Bayes	0.53	0.65
SI	3	Naive Bayes	0.56	0.65
SI	4	Naive Bayes	0.87	0.65
SI	5	Naive Bayes	0.74	0.65
SI	6	Naive Bayes	0.65	0.65
SI	1	Random Forest	0.64	0.70
SI	2	Random Forest	0.70	0.70
SI	3	Random Forest	0.66	0.70
SI	4	Random Forest	0.55	0.70
SI	5	Random Forest	0.74	0.70
SI	6	Random Forest	0.65	0.70
SI	1	Rede Neural	0.57	0.67
SI	2	Rede Neural	0.59	0.67
SI	3	Rede Neural	0.56	0.67
SI	4	Rede Neural	0.87	0.67
SI	5	Rede Neural	0.74	0.67
SI	6	Rede Neural	0.68	0.67
SI	1	SVM	0.62	0.70
SI	2	SVM	0.66	0.70
SI	3	SVM	0.52	0.70
SI	4	SVM	0.87	0.70
SI	5	SVM	0.74	0.70
SI	6	SVM	0.75	0.70

Tabela 13 – Análise de algoritmo por período CC

Curso	Período	Algoritmo	Acurácia	Acurácia Média
CC	1	Árvore de decisão	0.60	0.66
CC	2	Árvore de decisão	0.61	0.66
CC	3	Árvore de decisão	0.66	0.66
CC	4	Árvore de decisão	0.67	0.66
CC	5	Árvore de decisão	0.67	0.66
CC	6	Árvore de decisão	0.75	0.66
CC	1	KNN	0.54	0.57
CC	2	KNN	0.58	0.57
CC	3	KNN	0.60	0.57
CC	4	KNN	0.61	0.57
CC	5	KNN	0.53	0.57
CC	6	KNN	0.66	0.57
CC	1	Naive Bayes	0.56	0.63
CC	2	Naive Bayes	0.57	0.63
CC	3	Naive Bayes	0.60	0.63
CC	4	Naive Bayes	0.63	0.63
CC	5	Naive Bayes	0.66	0.63
CC	6	Naive Bayes	0.76	0.63
CC	1	Random Forest	0.61	0.64
CC	2	Random Forest	0.60	0.64
CC	3	Random Forest	0.66	0.64
CC	4	Random Forest	0.57	0.64
CC	5	Random Forest	0.66	0.64
CC	6	Random Forest	0.75	0.64
CC	1	Rede Neural	0.54	0.64
CC	2	Rede Neural	0.58	0.64
CC	3	Rede Neural	0.66	0.64
CC	4	Rede Neural	0.63	0.64
CC	5	Rede Neural	0.65	0.64
CC	6	Rede Neural	0.76	0.64
CC	1	SVM	0.56	0.66
CC	2	SVM	0.59	0.66
CC	3	SVM	0.66	0.66
CC	4	SVM	0.68	0.66
CC	5	SVM	0.66	0.66
CC	6	SVM	0.76	0.66

Tabela 14 – Análise de algoritmo por período EC

Curso	Período	Algoritmo	Acurácia	Acurácia Média
EC	1	Árvore de decisão	0.58	0.72
EC	2	Árvore de decisão	0.77	0.72
EC	3	Árvore de decisão	0.53	0.72
EC	4	Árvore de decisão	0.85	0.72
EC	5	Árvore de decisão	0.68	0.72
EC	6	Árvore de decisão	0.90	0.72
EC	1	KNN	0.55	0.63
EC	2	KNN	0.48	0.63
EC	3	KNN	0.61	0.63
EC	4	KNN	0.55	0.63
EC	5	KNN	0.69	0.63
EC	6	KNN	0.90	0.63
EC	1	Naive Bayes	0.55	0.69
EC	2	Naive Bayes	0.63	0.69
EC	3	Naive Bayes	0.57	0.69
EC	4	Naive Bayes	0.80	0.69
EC	5	Naive Bayes	0.65	0.69
EC	6	Naive Bayes	0.90	0.69
EC	1	Random Forest	0.57	0.70
EC	2	Random Forest	0.63	0.70
EC	3	Random Forest	0.59	0.70
EC	4	Random Forest	0.80	0.70
EC	5	Random Forest	0.69	0.70
EC	6	Random Forest	0.90	0.70
EC	1	Rede Neural	0.55	0.70
EC	2	Rede Neural	0.64	0.70
EC	3	Rede Neural	0.57	0.70
EC	4	Rede Neural	0.88	0.70
EC	5	Rede Neural	0.65	0.70
EC	6	Rede Neural	0.90	0.70
EC	1	SVM	0.58	0.70
EC	2	SVM	0.64	0.70
EC	3	SVM	0.53	0.70
EC	4	SVM	0.87	0.70
EC	5	SVM	0.68	0.70
EC	6	SVM	0.90	0.70

Após a análise dos cursos por semestre ficou constatado que o período com maior índice de acurácia para todos os algoritmos aqui elencados foram respectivamente o 4º período para os Cursos de Engenharia da Computação - EC e Sistemas de Informação - SI e o 6º período referente ao curso de Ciência da Computação - CC.

A diferença da acurácia tem relação direta com a quantidade de matérias presentes em cada semestre. O 4º período para SI tem somente 4 matérias, enquanto o curso de EC já tem uma base sólida de matérias exatas sendo o 4º período o último complemento. Para o curso de CC no 6º período é onde temos algumas matérias não tão complexas, o que contribui para evitar a evasão do curso. A partir desse resultado inicial constatamos que para os Cursos de Computação quase todos os algoritmos apresentam menor acurácia no 1º período. Já para os Cursos de Sistema de Informação este cenário se aplica ao 3º período, salvo em alguns algoritmos. No contexto de Engenharia da Computação alguns algoritmos apresentam sua baixa acurácia no 3º e outros no 4º período. Este pode ser um indicador para muitas matérias no mesmo período ou alta complexidade de diversas matérias em um único período. Este contexto influencia diretamente o desempenho do aluno e a acurácia do algoritmo.

4.3 Considerações Finais

Este estudo inicial apresentou a análise de diferentes técnicas de aprendizagem supervisionada no contexto da mineração de dados educacionais para prevenção do abandono escolar de estudantes universitários.

Enquanto alguns algoritmos apresentaram resultados sólidos para cada semestre (Florestas Aleatórias e Árvores de Decisão), alguns deles não foram capazes de atingir resultados tão elevados (ou seja, Redes Neurais), provavelmente porque precisam de mais dados ou uma etapa de ajuste de parâmetros. Assim sendo, o melhor algoritmo para o nosso estudo de caso é a Árvore de Decisão.

A partir de um algoritmo com um desempenho promissor em relação aos dados disponíveis é possível uma abordagem computacional mais assertiva do modelo preditivo. Assim se buscará a evolução do comportamento do algoritmo em relação aos atributos aqui estudados. A partir deste momento é possível propor uma abordagem para avaliar o desempenho de alunos em disciplinas. Esta abordagem deverá considerar a situação individual de cada instituição, o histórico das disciplinas, o departamento responsável pela disciplina em questão, o histórico do aluno e o currículo obrigatório que o aluno deve frequentar a cada semestre ou período.

De maneira a retratar a situação individual de cada instituição, é possível orientar a abordagem para avaliar o desempenho de alunos em disciplinas através de questões que retratem os principais problemas identificados em diferentes cenários de modo a mitigar a evasão do ensino superior.

5

Abordagem para avaliar o desempenho de alunos em disciplinas

O objetivo desta Seção 5 é apresentar a abordagem utilizada após os resultados iniciais desta pesquisa e seleção do algoritmo árvore de decisão para a continuação deste estudo. Aqui é observado a pesquisa descritiva, pois o foco será o estudo de caso, onde já foi obtido um pré-resultado descrito na Seção 4. O estudo de caso foi realizado subsequentemente a apresentação desta abordagem para analisar a eficiência e a viabilidade da abordagem proposta.

5.1 Objetivo da Abordagem

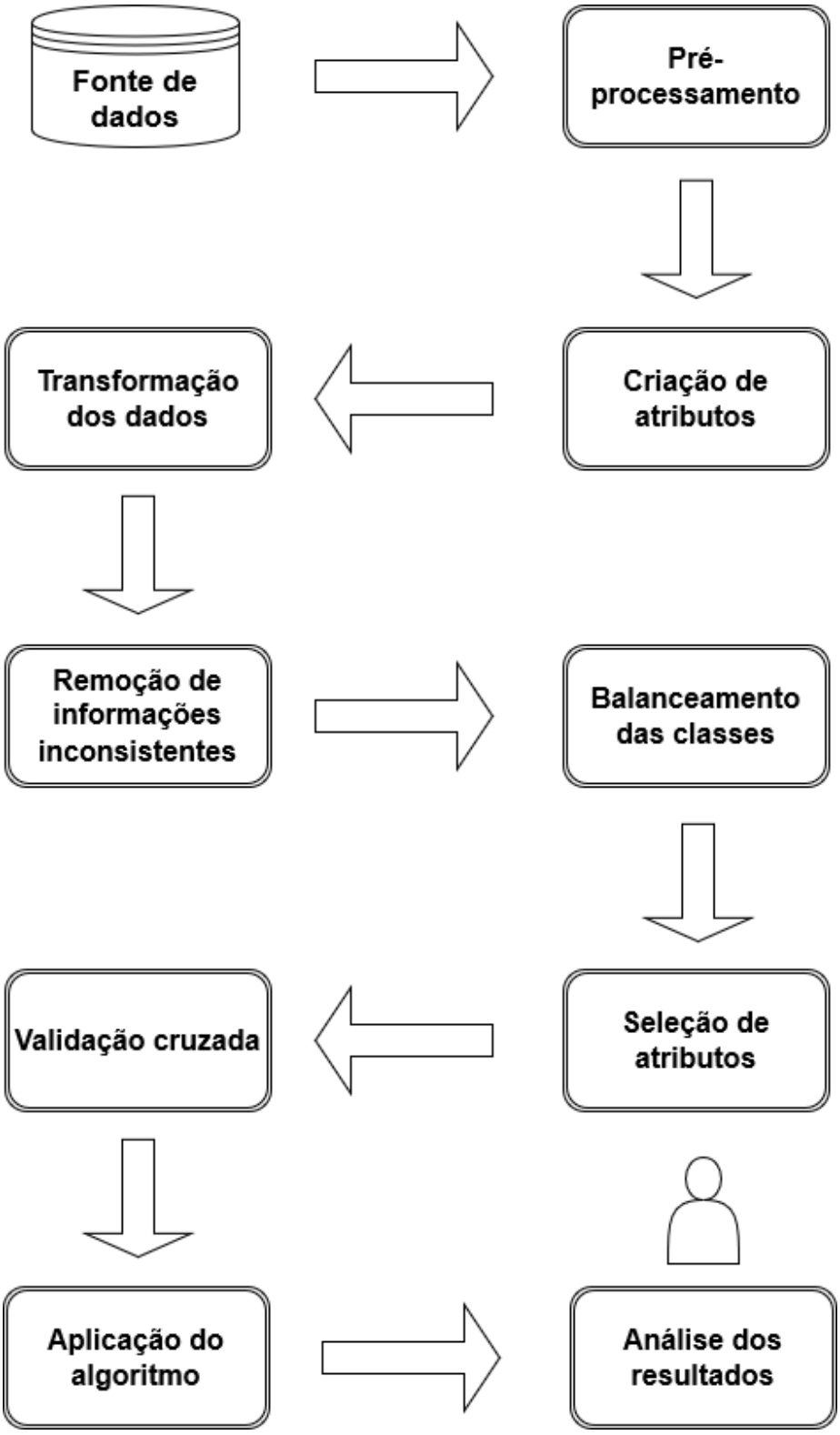
O que se deseja encontrar após a aplicação desta abordagem é uma maneira facilitada para interpretar as informações e assim permitir a análise de dados de forma intuitiva. Para que isto seja possível é necessário realizar várias transformações em atributos já existentes com o propósito de aumentar a acurácia do algoritmo árvore de decisão. Deste modo, o que a abordagem propõe é a criação de novos atributos a partir de outros já existentes.

Para que isso seja possível, é necessário realizar várias transformações nos dados existentes para criar estes novos atributos. Deste modo para analisar a abordagem proposta, em nosso estudo de caso, foram utilizados os dados oriundos da Universidade Federal de Sergipe em cursos do Departamento de Computação (DCOMP) entre os anos de 2007 a 2018. Assim sendo, as questões pesquisa formuladas foram direcionadas às dores do DCOMP e analisadas sob a ótica dos novos atributos criados.

5.2 Etapas na descoberta do conhecimento

De maneira geral, a Figura 19 mostra os principais pontos desenvolvidos até a construção do conhecimento para responder às questões propostas e facilitar a interpretação desses dados através da ferramenta KNIME.

Figura 19 – Etapas na descoberta do conhecimento



• Pré-processamento dos dados: Neste primeiro momento ocorrem as tarefas relacionadas a extração dos dados, limpeza de informações discrepantes, transformação da informação, carga e posteriormente a atualização destes dados processados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

• Criação de atributos: Esta proposta visa quantificar informações relevantes do conjunto apresentado de maneira eficaz e com significância maior que os atributos preexistentes. Este estudo apresenta a criação de novos atributos, que serão detalhados na Seção 5.3, considerando informações consolidadas na base de dados do Pentaho e pesquisa bibliográfica. A finalidade destes atributos recém criados é apresentar valores compreensíveis e de fácil interpretação. Estes novos valores ainda agregam indicativos para a tomada de decisão dos gestores educacionais, permitindo assim, ações direcionadas a alunos que tenham alto risco de evasão do curso selecionado.

• Transformação dos dados: Neste passo são aplicadas as regras de agregação, normalização, discretização e amostragem dos dados apresentados, técnicas estas, que já são empregadas tradicionalmente em trabalhos que envolvem mineração de dados segundo Han, Kamber e Pei (2011).

• Remoção de informações inconsistentes: Nesta etapa foram descartados valores discrepantes, como também, os dados incoerentes, registros que apresentavam quebra de relacionamento e remoção de tuplas que apresentavam várias colunas com elementos vazios.

• Balanceamento das classes: Neste cenário existe o desbalanceamento das classes que são o foco desta pesquisa. O desbalanceamento, por sua vez, pode enviesar a análise do algoritmo que passa a tendenciar classes mais frequentes e desconsiderar classes de frequência menor. Como consequência o classificador não apresenta resultado eficiente para instâncias de classes pouco representadas em nosso conjunto de dados (MARQUEZ-VERA; MORALES; SOTO, 2013). Uma das técnicas mais aplicadas para contornar esta situação é a utilização do algoritmo SMOTE (*Synthetic Minority Oversampling Technique*) para balancear as classes do subconjunto (CHAWLA et al., 2002). O algoritmo SMOTE, corrige a periodicidade das classes que apresentam frequência maior e menor, inserindo registros de classes minoritárias de acordo o algoritmo de agrupamento K-NN (AZUAJE, 2006).

• Seleção de Atributos: A seleção dos dados é uma técnica utilizada para a redução da dimensionalidade, onde são removidos os atributos que não possuem relevância, os que são redundantes e ainda os que são fracamente relevantes, conforme detalhado na Seção 2.4.2.

• Validação Cruzada: Após selecionar os atributos de maior relevância é necessário avaliar o desempenho do algoritmo que esta sendo utilizado. Nesta pesquisa, para mensurar esta performace foi adotada a médida acurácia. Durante os experimentos foi adotada a técnica de validação cruzada (fator k = 10), para a execução no conjunto selecionado.

• Aplicação do Algoritmo: Neste momento será aplicado o algoritmo de classificação com os atributos de maior relevância que foram selecionados. Caso surja a necessidade, pode-se criar outro *dataset* alternativo. No contexto da EDM é recomendado a utilização de algoritmos de classificação "caixa branca", que produzam modelos de fácil interpretação, e que podem ser empregados para a tomada de decisão (MARQUEZ-VERA; MORALES; SOTO, 2013). Para

os experimentos deste estudo foram avaliados no total 4.017 alunos do DCOMP desde o ano de 2007 até 2018 e 49.013 registros em turmas disponíveis para matrícula, composto de dados acadêmicos que serão detalhados a seguir.

5.3 Pré-processamento e Criação de Atributos

Para reproduzir a pesquisa apresentada, é necessário reproduzir as etapas já exemplificadas na Figura 19. Portanto, neste momento de pré-processamento é imprescindível que alguns atributos existam em seu banco de dados ou CSVs. Esses atributos serão necessários para a construção das novas variáveis a serem apresentadas neste trabalho. Esses atributos são exemplificados na Tabela 15.

Tabela 15 – Atributos básicos necessários para a criação de novos atributos.

Atributo	Novo Atributo
número de estudantes que foram aprovados na classe	atributo de dificuldade média da turma
número de estudantes que foram reprovados na classe	atributo de dificuldade média da turma
total de matérias em que o estudante foi aprovado	atributo de dificuldade média do aluno
total de matérias em que o estudante foi reprovado	atributo de dificuldade média do aluno

Estes atributos são a base fundamental para a pesquisa. Neste estudo de caso os atributos necessários para o cálculo do novo atributo de dificuldade média da turma estão localizados na tabela fttu_turma representadas através dos atributos fttu_aprovado e fttu_reprovado. A tabela fttu_turma reflete todas as informações referentes as matérias ofertadas aos estudantes. Em seguida, para o cálculo do novo atributo de dificuldade média do aluno é necessário um cruzamento entre as tabelas fttu_turma e ftda_desempenho_aluno, que é a tabela que concentra as informações referentes aos discentes da instituição. Ambas tabelas foram apresentadas na Seção 3.2 através da Figura 12.

A medida em que avançamos e transformamos esses dados, será possível criar e manipular novos atributos. A fim de reduzir a maldição da dimensionalidade, este trabalho propõe a criação de três novos atributos para auxiliar na tomada de decisão através do algoritmo árvore de decisão, e assim, obter maior acurácia e exatidão além de facilitar todo o processo de mineração de dados (MáRQUEZ et al., 2016), (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

5.3.1 Atributo de dificuldade média da turma

Para auxiliar na tomada de decisão pelo algoritmo, podemos criar um atributo que será responsável por avaliar a dificuldade média da disciplina diretamente relacionada à aula que será

cursada. Para esta pesquisa foi utilizada a fórmula abaixo:

$$DMT = \frac{AP + REP}{AP}$$

Fonte: (JÚNIOR et al., 2015) adaptado.

Onde:

$DMT \longrightarrow$ representa a dificuldade média da turma durante o período letivo;

$AP \longrightarrow$ quantidade de alunos que foram aprovados na turma;

$REP \longrightarrow$ quantidade de alunos que foram reprovados na turma;

Após o cálculo destes valores foi aplicado a técnica da normalização nos resultados obtidos para facilitar a interpretação por parte do algoritmo árvore de decisão. Com a determinação desta fórmula é possível então definir o "atributo de dificuldade média do aluno", uma vez que, os alunos podem apresentar desempenhos distintos entre si. Este atributo visa auxiliar a predição de evasão do aluno por turma, além de contribuir para a identificação de quais são as principais características de perfis de alunos evasivos do curso.

5.3.2 Atributo de dificuldade média do aluno

Para cálcular o atributo que representa a "dificuldade média do aluno", neste contexto, é necessário determinar a dificuldade média das turmas cursadas pelo mesmo durante a graduação, uma vez que, será analisado o desempenho do discente através da fórmula abaixo:

$$DMA = \frac{\sum_{i=1}^{n} DMTA - \sum_{i=1}^{n} DMTR}{A + R}$$

Fonte: (JÚNIOR et al., 2015) adaptado.

Onde:

$DMA \longrightarrow$ representa a dificuldade média das turmas cursadas pelo aluno;

$DMTA \longrightarrow$ representa a dificuldade média das turmas em que aluno foi aprovado;

$DMTR \longrightarrow$ representa a dificuldade média das turmas em que aluno foi reprovado;

$A \longrightarrow$ total de disciplinas em que o aluno obteve aprovação;

$R \longrightarrow$ total de disciplinas em que o aluno obteve reprovação;

5.3.3 Atributo de dificuldade média do período a ser cursado

Período, neste contexto, pode ser interpretado como o currículo de um semestre letivo, que pode variar em cada instituição. Para calcular o atributo que representa a ”dificuldade média do período”, é necessário determinar a ”dificuldade média das disciplinas” que compõem o período em que o aluno está matriculado. Este atributo visa medir a dificuldade dos semestres

dos cursos analisados neste estudo de caso. A fórmula é mostrada abaixo:

$$DMP = \frac{\sum_{i=1}^{n} DMT}{QTD}$$

Onde:

$DMT \longrightarrow$ representa as dificuldades médias das disciplinas que compõem o período escolar;

$QTD \longrightarrow$ representa o número de disciplinas obrigatórias que devem ser cursadas por semestre;

5.3.4 Transformação de dados e Remoção de inconsistências

Afim de reproduzir a pesquisa apresentada, sugere-se que existam duas tabelas principais, seja em um banco de dados relacional ou em arquivos CSVs distintos. A primeira tabela refere-se às aulas ou disciplinas oferecidas pela instituição, que neste estudo de caso é representado através da tabela fttu_turma. Neste ponto, é importante destacar o histórico de cada disciplina, pois essas informações serão essenciais para que seja possível determinar o valor de DMT ou simplesmente "Atributo de dificuldade média da turma".

Na tabela de turmas representada por fttu_turma usaremos a fórmula DMT já vista na Seção 5.3.1 para calcular o valor DMT da disciplina. Neste ponto, precisamos saber o número de alunos que foram aprovados e a quantidade de alunos que foram reprovados na classe. Neste estudo de caso os valores são representados através dos atributos fttu_aprovado e fttu_reprovado presentes na tabela fttu_turma. Caso não haja aprovações na disciplina analisada, o número representado será automaticamente o número de alunos reprovados. Quanto maior o valor DMT encontrado e quanto mais este valor é superior ao número 0, mais difícil é para o aluno realizar a disciplina. Após a realização desse cálculo, o valor encontrado em cada disciplina apresentada em cada período deve ser representado por meio de uma coluna que registrará o valor encontrado. Em nosso estudo este valor foi representado na coluna fttu_dificuldade_turma, coluna esta, recém-criada na tabela fttu_turma.

Ainda na tabela turmas, representada por fttu_turma, é necessário garantir que as informações a respeito da aprovação ou reprovação do aluno estejam presentes. Estes dados serão essenciais para que a árvore de decisão possa ter o treinamento correto a partir destas informações. Os dados podem ser representados por meio de um booleano, e nesta pesquisa, foi atribuído o valor 0 ou 1 para representar esta informação.

É importante avançar com a transformação de dados somente após obter o atributo DMT referente a todas as turmas ou disciplinas da instituição. Esta situação ocorre porque o atributo DMT está diretamente ligado ao atributo DMA, sendo este, fundamental para que seja possível

calcular o valor relacionado a dificuldade média do aluno a partir da dificuldade da turma a ser cursada.

Na tabela referente aos alunos, aqui representada por ftda_desempenho_aluno, usaremos a fórmula apresentada na Seção 5.3.2 para calcular o valor DMA ou simplesmente "Atributo de dificuldade média do aluno". Neste ponto, precisamos saber o número de disciplinas em que o aluno foi aprovado e o número de disciplinas em que o aluno foi reprovado. Em nosso estudo de caso estes valores podem ser obtidos através de um cruzamento entre as tabelas ftda_desempenho_aluno e fttu_turma. Em seguida, a partir da tabela de turmas (fttu_turma) será possível resgatar o valor DMT (atributo fttu_dificuldade_turma) correspondente às disciplinas em que o aluno foi aprovado e reprovado. Estes dados, após a aplicação da fórmula da Seção 5.3.2, devem ser representados por meio de uma coluna que registrará o valor encontrado na tabela fttu_turma. Neste estudo de caso o valor foi representado através do atributo fttu_dificuldade_aluno. Valores maiores que 0 significam que o aluno está tendo um bom desempenho, enquanto valores abaixo de 0 demonstram que o aluno requer atenção.

Se necessário, dependendo do contexto de cada instituição e das questões colocadas, é possível criar uma nova tabela que terá a função de recolher as informações relacionadas com os valores de DMA dos alunos, DMT das turmas e o novo atributo DMP.

O atributo DMP representa a dificuldade do semestre de um curso específico em uma instituição. Para calcular o DMP é necessário já ter os valores de DMT das classes. Em nosso estudo de caso, essas informações serão utilizadas para observar como o currículo obrigatório está estruturado e como esta estrutura pode influenciar a evasão dos alunos.

Para que esses dados sejam melhor assimilados pela árvore de decisão, os valores DMA e DMT devem ser normalizados usando a técnica Min-Max apresentada na Seção 2. Estes valores após a normalização devem passar por uma transformação disponível no ambiente KNIME chamada *Auto-Binner*. Este recurso permite agrupar dados numéricos em intervalos - chamados *bins*. Nesta pesquisa, um número fixo de *bins* igual a 5 foi usado. Então, se houver campos com valores numéricos, estes devem ser transformados em categóricos.

Após as transformações realizadas, alguns dados podem apresentar registros duplos, inconsistências ou mesmo a falta dos atributos de base necessários para as transformações e criações dos novos atributos apresentados. Especialmente se os dados são oriundos de arquivos CSVs. Estes dados inconsistentes podem ser removidos para avançar.

5.3.5 Seleção de Atributos, Balanceamento de classes e Validação cruzada

Após essas etapas, usaremos o Knime para auxiliar no processamento desses dados. Inicialmente iremos carregar todos os registros da tabela ou CSV para a classe. Posteriormente utilizaremos a funcionalidade *filter* para selecionar as três colunas que representam o valor DMT, o valor DMA e o campo referente à aprovação ou reprovação do aluno.

O ambiente Knime funciona através de nós. Neste experimento o nosso conjunto de dados foi particionado em 70-30, ou seja, 70% dos registros foram usados para treinar o classificador e 30% foram utilizados para testar o modelo preditivo gerado por meio da árvore de decisão. O algoritmo SMOTE foi aplicado com o objetivo de balancear o conjunto de dados em nó subsequente.

Neste ponto, para validar o desempenho do algoritmo a técnica validação cruzada foi aplicada em todo *dataset* utilizando o fator k=10 conforme apresentado na Seção 2. Neste momento, foi utilizado o nó responsável pela validação cruzada, que também já aplica o partionamento no conjunto de dados em *subdatasets*.

5.3.6 Aplicação do algoritmo e Análise dos resultados

Em nosso estudo de caso, a árvore de decisão foi aplicada após todas estas etapas. É importante notar que nesta abordagem a matriz de confusão e a acurácia foram utilizadas para avaliar os resultados.

Assim sendo, é possível responder a diferentes questões em diferentes cenários ou contextos de diferentes instituições. Em geral, a árvore de decisão obteve 98% de acerto, mas a principal contribuição é como usaremos os dados para responder e identificar pontos que ajudem os alunos, e assim, evitar a evasão dos discentes.

Neste estudo de caso, a abordagem apresentada foi aplicada no Departamento de Computação da UFS com o intuito de responder às seguintes questões de pesquisa:

QP1. *Quais matérias de grade curricular retém os alunos e consequentemente elevam a evasão?*

O objetivo é analisar o número de reprovações dos alunos, o histórico de reprovações das disciplinas e qual o departamento responsável por cada aula. Para isso, é necessário identificar as disciplinas com mais reprovações entre os alunos, identificar quais departamentos são responsáveis por essas disciplinas e avaliar por meio do atributo DMT qual a dificuldade dessas disciplinas.

Neste contexto será aplicado o algoritmo árvore de decisão e todas as etapas apresentadas na Figura 19 serão reproduzidas. Para esta abordagem, o atributo DMT não será normalizado. É esperado que a árvore de decisão mantenha a acurácia apresentada e utilizaremos os recursos de ordenação do Knime para explorar as turmas que apresentam maior índice de reprovação através de altos valores refletidos no atributo DMT das turmas analisadas. As disciplinas encontradas serão confrontadas com o relatório gerado através do Pentaho para validação da informação gerada através do algoritmo EDM.

É importante destacar que o histórico destas disciplinas pode ser revisada sob a percepção do novo atributo DMT. O valor DMT 1 indica que todos os alunos matriculados na

disciplina foram aprovados e, em contrapartida, o valor DMT 30 representa que todos os alunos matriculados na disciplina foram reprovados.

QP2. *Quais são as principais características de estudantes que largarão o curso selecionado?*

Nesta etapa será aplicado o algoritmo árvore de decisão em dois grupos distintos de alunos. Em um conjunto de dados somente teremos alunos aprovados e no segundo conjunto somente alunos reprovados. Para esta abordagem todas as etapas apresentadas na Figura 19 serão reproduzidas e será aplicada a técnica de padronização Min-Max [0,0, 1,0] para que os valores assumidos por DMA e DMT respeitem a dimensão estabelecida entre 0 e 1.

Como saída teremos 2 CSVs distintos, cada um referente a cada conjunto de dados respectivamente. Por meio dos novos atributos propostos, se analisa o padrão dos alunos que possuem melhor rendimento e dos que não alcançam tal aproveitamento no ensino superior. Neste contexto, pretendemos analisar os alunos classificados como reprovados, bem como, os alunos classificados como aprovados. Para que esta análise seja possível, os atributos a serem observados serão os novos atributos DMT e DMA ambos definidos na Seção 5.3.

No contexto de alunos repetentes, o atributo DMT com o valor 0 determina que nos registros destes alunos a nota não é fator determinante para reprovação, sendo categorizada como faltas e outros motivos característicos. Neste momento, quanto maior a proximidade do atributo DMT do valor 1, pior é a situação do(s) aluno(s). O atributo DMT com valor 1 indica que todos os alunos matriculados na disciplina foram reprovados. Para aulas com valor de atributo DMT que flutua entre 0 e 1, indica que houve aprovação e reprovação na disciplina analisada.

No contexto de alunos aprovados, o atributo DMT com valor 0 indica que houve aprovação total da turma, ou seja, todos os alunos inscritos foram aprovados. É importante destacar que o atributo DMA indicará como os alunos estão concentrados nos dois cenários, seja reprovação ou aprovação. As respostas encontradas em **QP1.** Podem complementar as informações descobertas durante a avaliação destes atributos durante **QP2.**

QP3. *A quantidade de matérias selecionadas influencia o desempenho do aluno e do algoritmo que está sendo aplicado?*

Para esta questão, será aplicado o algoritmo árvore de decisão em um determinado conjunto de dados de alunos. Este conjunto de dados é composto por alunos com matrícula ativa e que foram aprovados em disciplinas ofertadas entre o 1º ao 3º período. Este conjunto de dados possui alunos oriundos dos cursos de Sistemas de Informação, Ciência da Computação e Engenharia da Computação entre os anos de 2007 a 2018. Para prosseguir com a análise, todas os passos apresentados na Figura 19 serão reproduzidos.

Para auxiliar o cruzamento de informações, dentro do Knime foi criado um repositório

com a grade obrigatória de cada curso do Dcomp. O que permite contabilizar o número de disciplinas do 1º ao 3º período obrigatório de cada curso analisado. Dessa forma, cada curso e cada período foi isolado respeitando os componentes da grade curricular obrigatória que compõem os semestres letivos.

Após o cruzamento destas informações, como saída, teremos um arquivo CSV em que as aprovações de cada subconjunto podem ser observadas de acordo com os componentes curriculares obrigatórios. É esperado que o algoritmo ávore de decisão mantenha a taxa de acerto e que não apresente variação brusca em sua acurácia. As respostas encontradas em **QP1.** e **QP2.** podem complementar as informações encontradas.

QP4. *A correlação de matérias por semestre influencia o desempenho do aluno?*

Para esta questão, será aplicado o algoritmo árvore de decisão em um determinado conjunto de dados que possui alunos oriundos dos cursos de Sistemas de Informação, Ciência da Computação e Engenharia da Computação entre os anos de 2007 a 2018. Para esta análise todas as etapas apresentadas na Figura 19 serão reproduzidas, no entanto, o atributo DMT não será normalizado.

Nesta etapa, como não é aplicada a técnica de normalização Min-Max [0.0,1.0] no atributo DMT, o valor deste seguirá o contexto de variação do número 1 ao 30. O valor DMT 1 indica que todos os alunos matriculados na turma foram aprovados e, em contrapartida, o DMT 30 indica que todos os alunos matriculados na disciplina foram reprovados.

Para auxiliar o cruzamento de informações, será reutilizado o repositório criado dentro do Knime com a grade obrigatória de cada curso do Dcomp já citado em **QP3.** Para responder a esta questão pode-se analisar o histórico das disciplinas obrigatórias do 1º ao 3º período de cada curso. A análise levará em consideração o atributo DMT que não foi normalizado.

Neste ponto será obtido o valor DMT médio de todas as matérias já ofertadas, por exemplo, a disciplina Cálculo I teve mais de 800 turmas ofertadas entre os anos de 2007 a 2018. Neste raciocínio, os valores DMT de todas as disciplinas Cálculo I são somados e posteriormente será obtido o valor médio de todos DMT apresentados. Este valor DMP médio representará a dificuldade média de turmas de Cálculo I do Dcomp.

Posteriormente ao cruzamento das informações, a partir do DMP médio e a grade curricular obrigatória é calculado o novo atributo DMP citado na Seção 5.3. Este corresponde ao grau de dificuldade média do período a ser cursado em relação ao período letivo obrigatório. Como saída teremos um arquivo CSV em que cada curso do Dcomp pode ser observado de acordo com os valores DMP correspondentes aos seus componentes curriculares obrigatórios.

QP5. *Quais são os motivos que impulsionam a evasão dos alunos de cursos de computação?*

Todas as respostas encontradas nas questões anteriores são refletidas como os principais pontos de evasão. Essas informações são reunidas de forma que sejam destacadas e sejam possíveis intervenções para prevenir o abandono escolar.

Em geral, EDM nos permite fazer alterações quando necessário (SILVA, 2015). Assim, para obter os resultados esperados em diferentes cenários, podemos modificar os passos adotados. Então poderemos avaliar os resultados encontrados em diferentes estudos experimentais.

5.4 Resultados do desempenho de alunos em disciplinas

Para avaliar a abordagem proposta, foi realizado um estudo de caso com dados de alunos da Universidade Federal de Sergipe em cursos do Departamento de Informática (DCOMP). Assim, foi possível determinar o objeto de estudo e observar os efeitos que cada variável produz no objeto. Os atributos básicos mencionados na Seção 5.3 foram explorados exatamente conforme descrito e, além da abordagem computacional, também foi realizada pesquisa sobre o histórico das disciplinas, o departamento ao qual essas disciplinas obrigatórias estão vinculadas, as aprovações e reprovações dos alunos. Alguns departamentos relacionados aos componentes do currículo obrigatório também serão mencionados neste estudo de caso. Para um melhor contexto, os resultados encontrados são segmentados em abordagem computacional e por questão apresentada.

5.4.1 Abordagem Computacional

Para tanto, foram avaliados 4.017 dados de alunos DCOMP de 2007 a 2018 para os experimentos deste estudo e 49.013 registros em turmas disponíveis para inscrição, a partir de dados acadêmicos de arquivos CSVs da plataforma SIGAA da Universidade Federal de Sergipe (I.; ROSA, 2013).

5.4.1.1 Pré-processamento e Criação de atributos

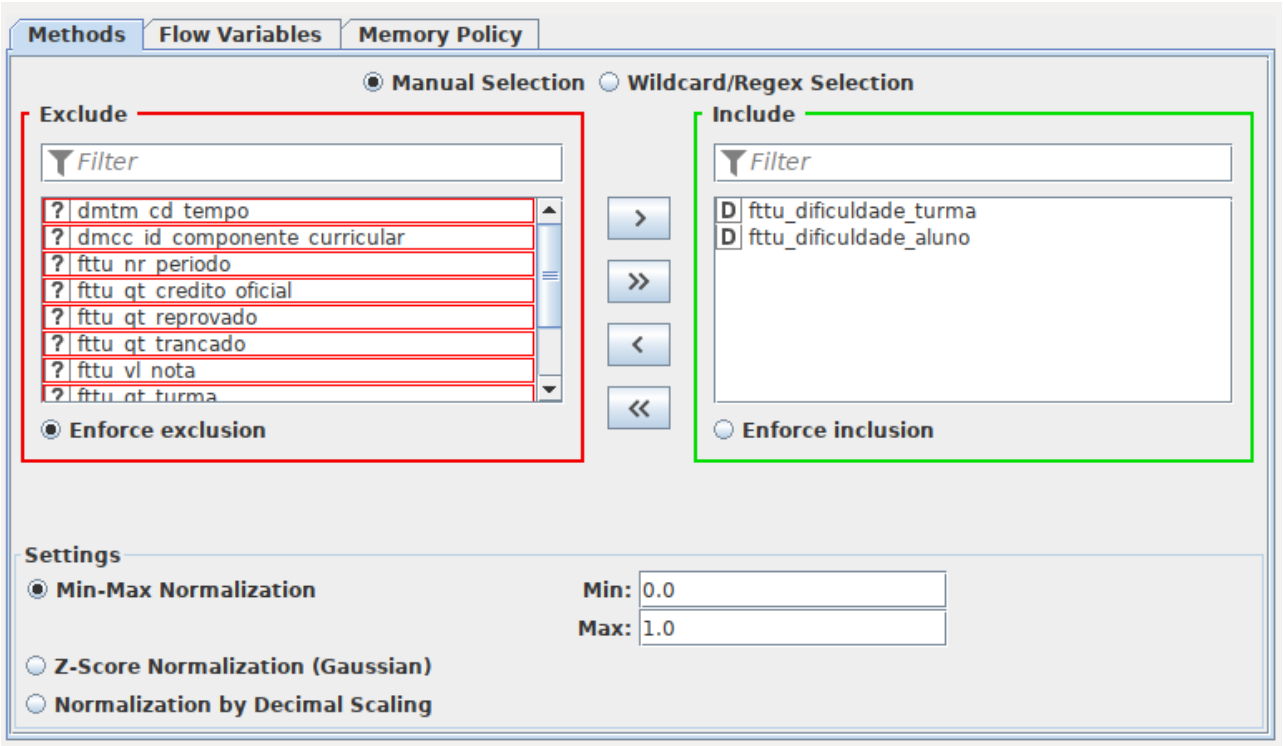
Neste momento, de acordo com a Seção 5.3, duas tabelas principais foram selecionadas. Uma desta com apenas os dados de todos os alunos e outra com os dados de todas as turmas oferecidas. Os arquivos já continham os atributos básicos necessários para a evolução do estudo de caso, bem como a criação dos novos atributos propostos.

5.4.1.2 Transformação de dados e Remoção de informações inconsistentes

Em seguida, as transformações dos dados foram realizadas para que fosse possível obter os atributos DMT e DMA. Os valores DMP também foram obtidos, mas este atributo será utilizado apenas para auxiliar o entendimento das questões levantadas em **QP4.** Os valores DMA

e DMT devem ser normalizados utilizando a técnica Min-Max e o recurso *binned* em ambiente KNIME através da funcionalidade *Auto-Binner*. Registros duplos, inconsistentes ou nulos foram removidos.

Figura 20 – Normalização no ambiente Knime.



5.4.1.3 Seleção de Atributos, Balanceamento de Classes e Validação Cruzada

Com o carregamento de todos os registros da tabela ou CSV para o ambiente KNIME. Usamos a funcionalidade *filter* para selecionar as quatro colunas que representam o valor DMT, o valor DMA, o crédito real do discente e o campo que se refere à aprovação ou reprovação do aluno para uma única tabela. Os valores categorizados por meio do KNIME *Auto-Binner* também são agregados nesta nova formação.

Figura 21 – Atributos selecionados.

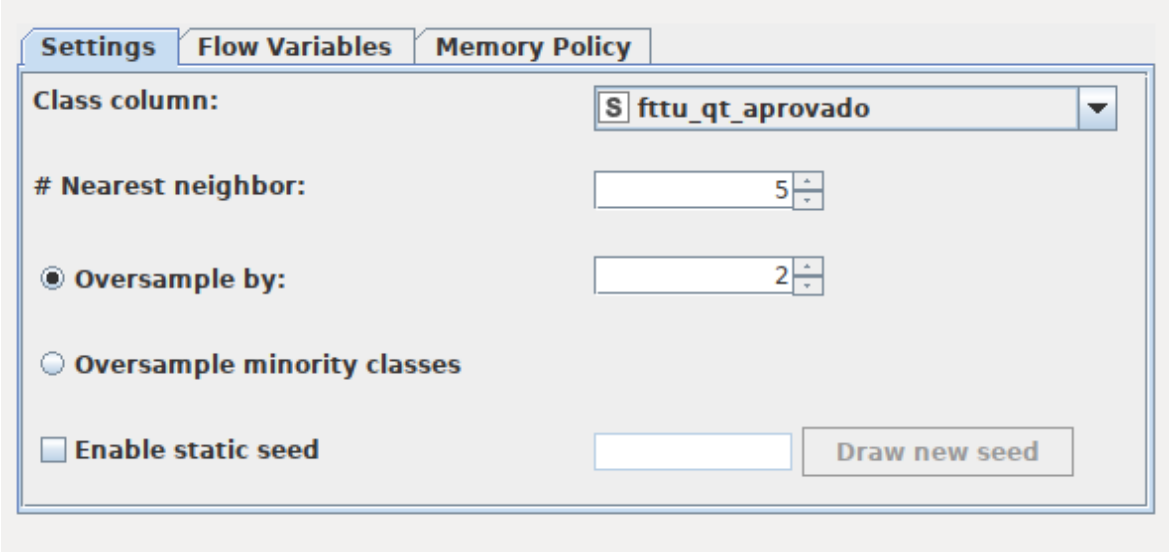
File Hilite Navigation View

Table "default" - Rows: 49013 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	fttu_qt_credito_real	fttu_qt_aprovado	fttu_dificuldade_turma	fttu_dificuldade_aluno	fttu_qt_credito_real [Binned]	fttu_dificuldade_turma [Binned]	fttu_dificuldade_aluno [Binned]
Row1	0	0	1	0.462	Bin 1	Bin 5	Bin 3
Row2	0	0	1	0.332	Bin 1	Bin 5	Bin 2
Row3	0	0	1	0.4	Bin 1	Bin 5	Bin 2
Row4	0	0	1	0.233	Bin 1	Bin 5	Bin 2
Row5	0	0	1	0.201	Bin 1	Bin 5	Bin 2
Row6	0	0	1	0.4	Bin 1	Bin 5	Bin 2
Row7	0	0	1	0.182	Bin 1	Bin 5	Bin 1
Row8	0	0	0.103	0.572	Bin 1	Bin 1	Bin 3
Row9	0	0	1	0.421	Bin 1	Bin 5	Bin 3
Row10	0	0	1	0.232	Bin 1	Bin 5	Bin 2
Row11	4	1	0.138	0.897	Bin 4	Bin 1	Bin 5
Row12	0	0	1	0.4	Bin 1	Bin 5	Bin 2
Row13	0	0	1	0.232	Bin 1	Bin 5	Bin 2

O conjunto de dados foi dividido em 70-30 e o algoritmo SMOTE foi aplicado. A técnica de validação cruzada também foi aplicada usando o fator k=10.

Figura 22 – Valores assumidos para o algoritmo SMOTE.



5.4.1.4 Aplicação do algoritmo e Análise de resultados

Com essa seleção, 14.704 registros foram avaliados por meio do modelo construído. Desses registros, 14.505 foram classificados corretamente e apenas 199 foram classificados incorretamente. A precisão alcançada pelo modelo de identificador de perfil evasivo apresentou 98,647% de acurácia, conforme detalhado na Figura 23, por meio da matriz de confusão emitida pelo KNIME.

Figura 23 – Acurácia alcançada através do modelo treinado.

fttu_qt_aprovado \ Prediction (fttu_qt_aprovado)	0	1
0	7438	90
1	109	7067

Correct classified: 14.505 **Wrong classified: 199**

Accuracy: 98,647 % **Error: 1,353 %**

Cohen's kappa (κ) 0,973

Após o algoritmo EDM alcançar a acurácia de 98,647%, o modelo classificador foi submetido a técnica de validação cruzada (fator n = 10). Como resultado, o *dataset* foi dividido em 10 *subdatasets* independentes com registros randômicos selecionados oriundos do *dataset* original. Após cada *subdataset* ter seu conjunto de dados predefinido, em seguida, foi aplicado o algoritmo SMOTE para balancear cada *subdataset* previamente a análise do classificador em cada subconjunto de dados. A Figura 24 demonstra a configuração utilizada para a validação cruzada no ambiente KNIME, bem como a Figura 25, define a estrutura de nós para a aplicação tanto do algoritmo SMOTE quanto o modelo preditivo treinado.

Figura 24 – Definição de parâmetros para a validação cruzada.

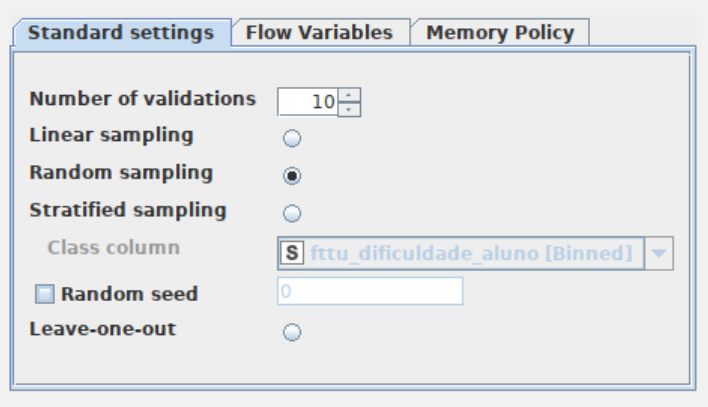
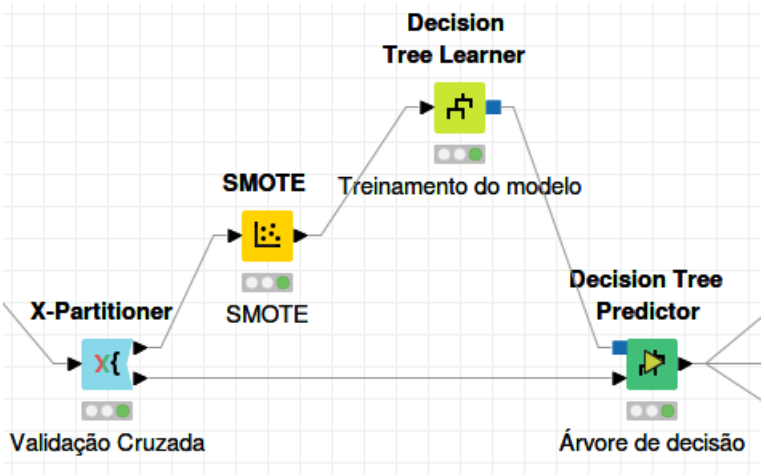


Figura 25 – Estrutura de nós para balancear e classificar *subdatasets*.



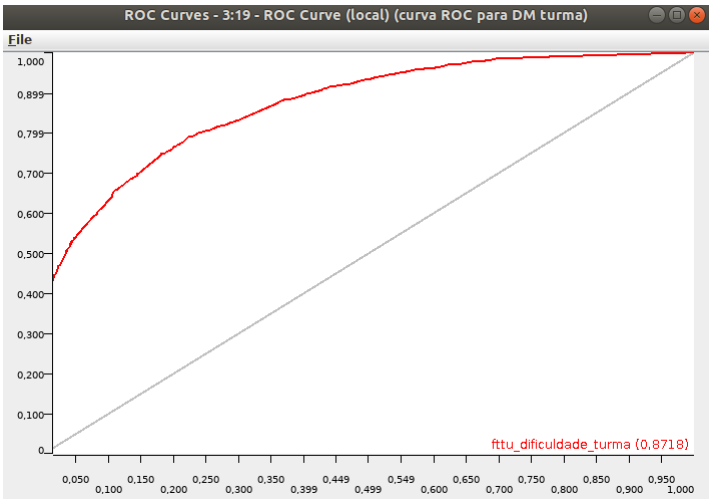
Com a utilização da validação cruzada foi possível analisar o comportamento de 10 *subdatasets* isolados entre si. A acurácia de cada subconjunto se manteve estável e apresentou 98,898% de aproveitamento em *subdataset* específico. A Tabela 16 apresenta a porcentagem de erro, o tamanho do *subdataset*, acurácia de cada subconjunto e a quantidade de erros encontrados durante a avaliação do modelo preditivo.

Tabela 16 – Valores obtidos em cada *Subdataset*.

Subdatasets	Erro em %	Tamanho do *Subdataset*	Acurácia em %	Erros do algoritmo
fold 0	1.510	4902	98.490	74
fold 1	1.163	4901	98.837	57
fold 2	1.326	4901	98.674	65
fold 3	1.489	4902	98.511	73
fold 4	1.224	4901	98.776	60
fold 5	1.102	4901	98.898	54
fold 6	1.306	4902	98.694	64
fold 7	1.306	4901	98.694	64
fold 8	1.489	4901	98.511	73
fold 9	1.469	4901	98.531	72

Paralelamente, a análise ROC - *Receiver Operating Characteristic* também foi aplicada nesta pesquisa. A Figura 26 mostra a curva ROC para avaliação do atributo criado DMT cujo resultado atingiu o valor de 0,8718.

Figura 26 – Análise ROC para o atributo DMT.



Para a tomada de decisão por parte do docente ou gestor instituicional, o KNIME possibilita o filtro dos alunos através do código da disciplina a que se deseja analisar. Deste modo é possível carregar todos os alunos do semestre letivo por disciplina, inclusive as matérias lecionadas pelo próprio professor. Uma vez que o modelo já esta devidamente treinado, o algoritmo árvore de decisão replicará a predição aplicada aos alunos devidamente cadastrados na(s) turma(s) e poderá exportar os resultados para um arquivo CSV.

Figura 27 – Seleção de disciplinas em que se deseja predizer a evasão/desistência dos alunos.

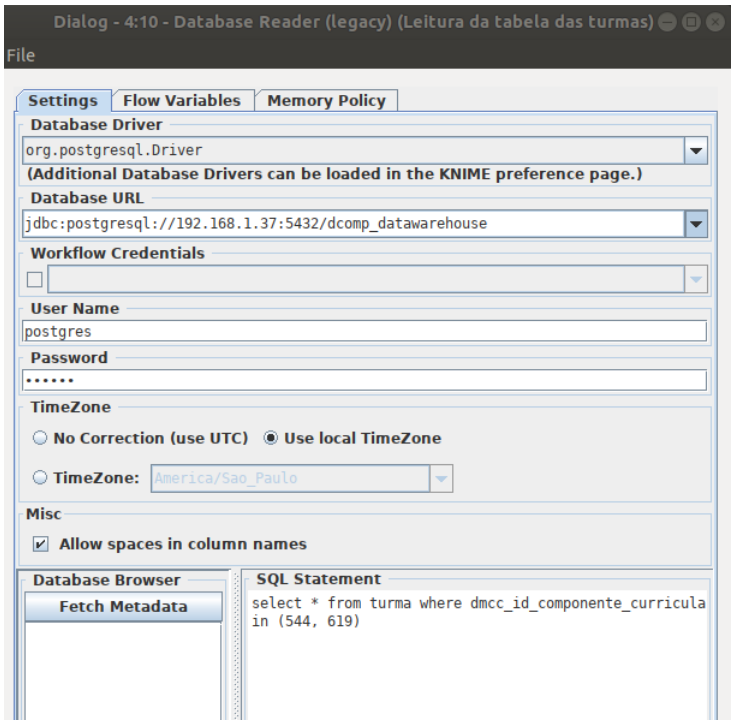
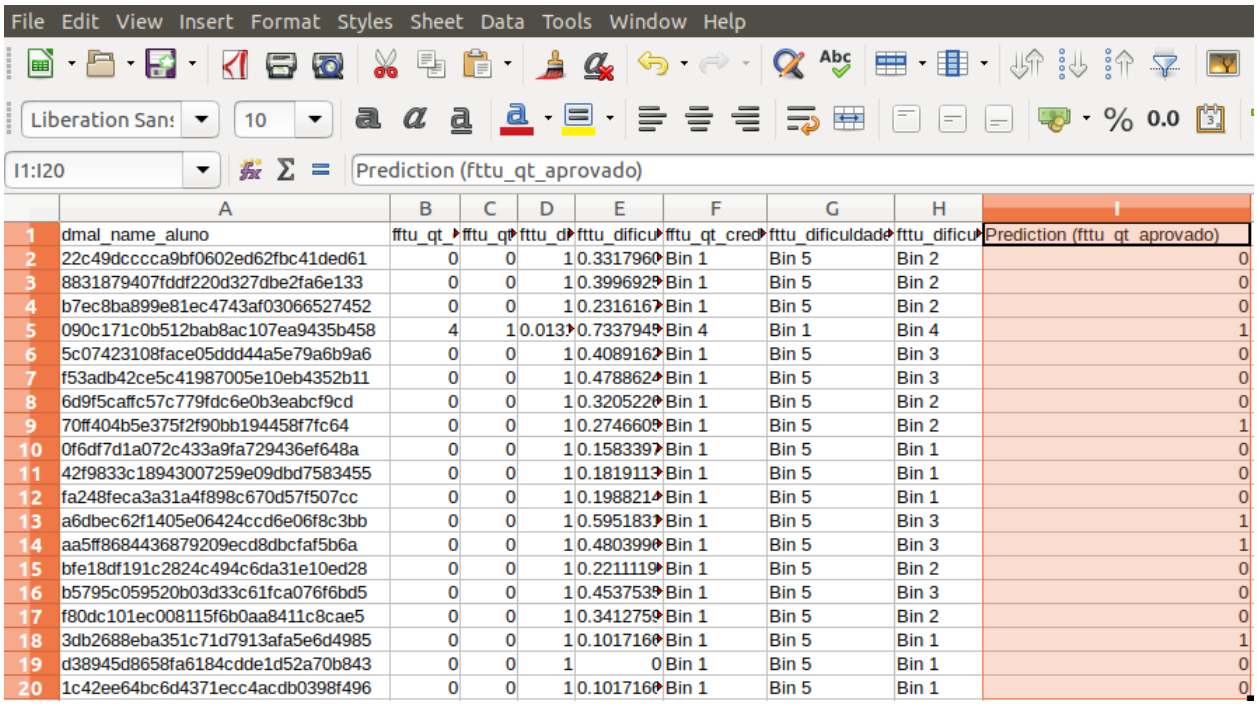


Figura 28 – Predição dos alunos que podem perder a disciplina ainda a ser cursada.



File Edit View Insert Format Styles Sheet Data Tools Window Help

Liberation San: 10

I1:I20 Prediction (fttu_qt_aprovado)

	A	B	C	D	E	F	G	H	I
1	dmal_name_aluno	fttu_qt_	fttu_qt	fttu_di	fttu_dificu	fttu_qt_cred	fttu_dificuldade	fttu_dificu	Prediction (fttu_qt_aprovado)
2	22c49dcccca9bf0602ed62fbc41ded61	0	0	1	0.3317960	Bin 1	Bin 5	Bin 2	0
3	8831879407fddf220d327dbe2fa6e133	0	0	1	0.3996929	Bin 1	Bin 5	Bin 2	0
4	b7ec8ba899e81ec4743af03066527452	0	0	1	0.2316167	Bin 1	Bin 5	Bin 2	0
5	090c171c0b512bab8ac107ea9435b458	4	1	0.013	0.7337949	Bin 4	Bin 1	Bin 4	1
6	5c07423108face05ddd44a5e79a6b9a6	0	0	1	0.4089162	Bin 1	Bin 5	Bin 3	0
7	f53adb42ce5c41987005e10eb4352b11	0	0	1	0.4788624	Bin 1	Bin 5	Bin 3	0
8	6d9f5caffc57c779fdc6e0b3eabcf9cd	0	0	1	0.3205220	Bin 1	Bin 5	Bin 2	0
9	70ff404b5e375f2f90bb194458f7fc64	0	0	1	0.2746609	Bin 1	Bin 5	Bin 2	1
10	0f6df7d1a072c433a9fa729436ef648a	0	0	1	0.1583397	Bin 1	Bin 5	Bin 1	0
11	42f9833c18943007259e09dbd7583455	0	0	1	0.1819113	Bin 1	Bin 5	Bin 1	0
12	fa248feca3a31a4f898c670d57f507cc	0	0	1	0.1988214	Bin 1	Bin 5	Bin 1	0
13	a6dbec62f1405e06424ccd6e06f8c3bb	0	0	1	0.5951837	Bin 1	Bin 5	Bin 3	1
14	aa5ff8684436879209ecd8dbcfaf5b6a	0	0	1	0.4803990	Bin 1	Bin 5	Bin 3	1
15	bfe18df191c2824c494c6da31e10ed28	0	0	1	0.2211119	Bin 1	Bin 5	Bin 2	0
16	b5795c059520b03d33c61fca076f6bd5	0	0	1	0.4537539	Bin 1	Bin 5	Bin 3	0
17	f80dc101ec008115f6b0aa8411c8cae5	0	0	1	0.3412759	Bin 1	Bin 5	Bin 2	0
18	3db2688eba351c71d7913afa5e6d4985	0	0	1	0.1017160	Bin 1	Bin 5	Bin 1	1
19	d38945d8658fa6184cdde1d52a70b843	0	0	1	0	Bin 1	Bin 5	Bin 1	0
20	1c42ee64bc6d4371ecc4acdb0398f496	0	0	1	0.1017160	Bin 1	Bin 5	Bin 1	0

5.4.2 Quais matérias de grade curricular retém os alunos e consequentemente elevam a evasão?

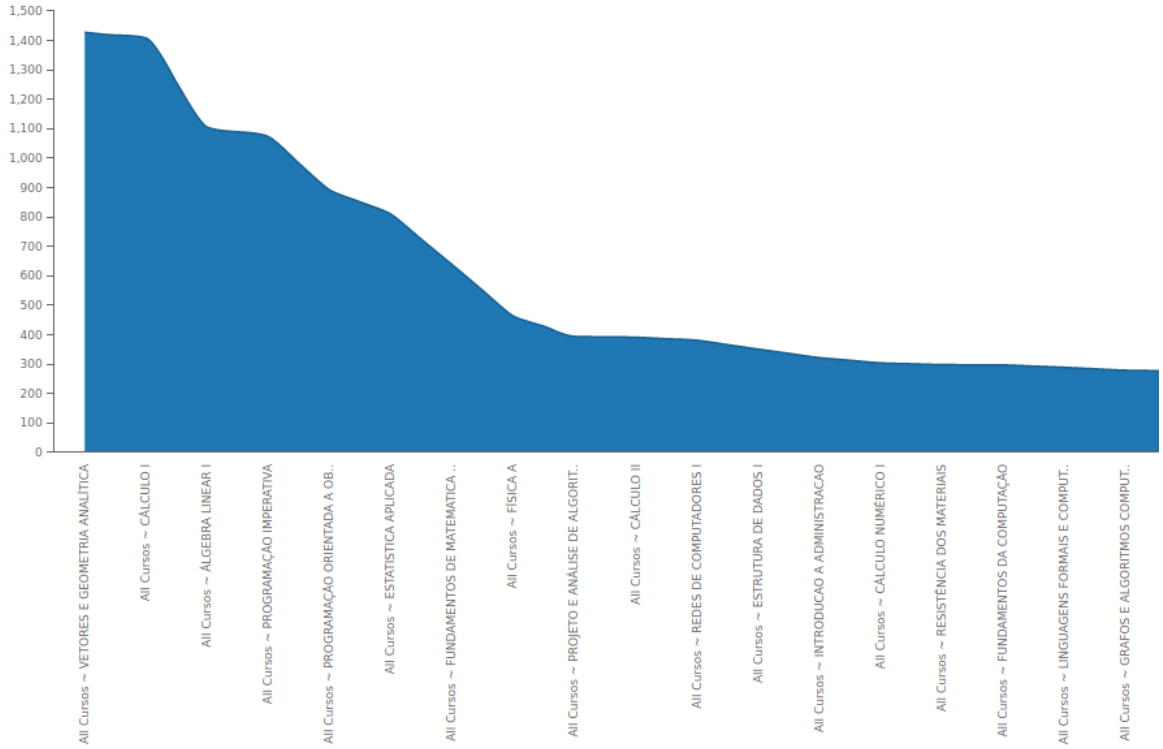
Para responder esta questão os dados foram classificados e reorganizados através de *Report* no ambiente DW Pentaho, que retrata quais são as disciplinas que apresentam maior quantidade de reprovação durante o intervalo de 2007 até 2018 de todos os cursos do Dcomp - UFS. A Figura 29 lista as dez disciplinas que mais reprovam em cursos ofertados pelo Dcomp e a Figura 30 demonstra a distribuição das reprovações destas.

Figura 29 – Lista de disciplinas que obtiveram mais reprovações entre 2007 até 2018.

Disciplinas_Com_Mais_Reprovacao.saiku

		All Períodos
(All)	Componente curricular	Reprovados
All Cursos	VETORES E GEOMETRIA ANALÍTICA	1,427
	CÁLCULO I	1,408
	ÁLGEBRA LINEAR I	1,105
	PROGRAMAÇÃO IMPERATIVA	1,073
	PROGRAMAÇÃO ORIENTADA A OBJETOS	892
	ESTATISTICA APLICADA	810
	FUNDAMENTOS DE MATEMATICA PARA COMPUTACAO	639
	FÍSICA A	464
	PROJETO E ANÁLISE DE ALGORITMOS	394
	CÁLCULO II	391

Figura 30 – Distribuição de matérias com maior índice de reprovação.



Para auxiliar a análise, as matérias apresentadas na Figura 29 e 30 foram segmentadas por departamento origem, e também é considerado o histórico das disciplinas apresentadas através do atributo DMT definido na Seção 5.3. Este atributo apresenta um contexto que pode variar do número 1 ao 30. O número 1 indica que todos os alunos matriculados na turma foram aprovados, e em contraparte, o número 30 indica que todos os alunos registrados na disciplina foram reprovados. A Tabela 17 mostra os departamentos correspondentes, a Tabela 18 retrata as aprovações/reprovações sob o atributo DMT e a Tabela 19 retrata porcentagem destas.

Tabela 17 – Departamento origem e disciplinas relacionadas.

Departamento Origem	Disciplina	Sigla da Disciplina
DECAT	Estatística Aplicada	EA
DCOMP	Programação Imperativa	PI
DCOMP	Programação Orientada a Objetos	POO
DCOMP	Projeto e Análise de Algoritmos	PAA
DFI	Física A	FA
DMA	Vetores e Geometria Analítica	VGA
DMA	Cálculo I	CI
DMA	Álgebra Linear I	ALI
DMA	Fundamentos de Matemática para Computação	FMC
DMA	Cálculo II	CII

A ementa dos cursos de computação é composta por outras disciplinas além das já ofertadas pelo Dcomp. Este conteúdo pode ser ofertado por outros Departamentos como, por exemplo,

o Departamento de Matemática - DMA, o Departamento de Física - DFI, o Departamento de Estatística e Ciências Atuarias - DECAT, entre outros. Dentre a lista que elenca as dez disciplinas que mais reprovam em cursos ofertados pelo Dcomp é possível constatar que 50% destas fazem parte do Departamento de Matemática - DMA, em seguida destaca-se que 30% das disciplinas são oriundas do próprio Departamento de Computação, em seguida o Departamento de Física - DFI e o Departamento de Estatística e Ciências Atuarias - DECAT representam 10% da lista cada um respectivamente.

Tabela 18 – Histórico das disciplinas ofertadas entre 2007 até 2018.

Sigla	Turmas Ofertadas	Turmas - DMT 1	Turmas - DMT 2 a 29	Turmas - DMT 30
EA	488	151	160	177
PI	200	13	95	92
POO	138	5	82	51
PAA	30	1	17	12
FA	237	29	74	134
VGA	522	55	106	361
CI	508	46	108	354
ALI	417	49	138	230
FMC	52	4	32	16
CII	214	42	51	121

Tabela 19 – Histórico das disciplinas em porcentagem entre 2007 até 2018.

Sigla	Turmas Ofertadas	100% Aprovadas	Turmas Parciais	100% Reprovadas
EA	488	30.94%	32.78%	36.27%
PI	200	6.5%	47.5%	46%
POO	138	3.62%	59.42%	36.95%
PAA	30	3.33%	56.66%	40%
FA	237	12.23%	31.22%	56.54%
VGA	522	10.53%	20.30%	69.15%
CI	508	9.05%	21.25%	69.68%
ALI	417	11.75%	33.09%	55.15%
FMC	52	7.69%	61.53%	30.76%
CII	214	14.95%	23.83%	56.54%

Este resultado retrata que 70% das disciplinas que mais retém os alunos do Dcomp não são exclusivas do Departamento de Computação. As disciplinas que pertecem ao Departamento de Matemática apresentam altos índices de reprovação conforme demonstrado na Tabela 19. Dentre as 5 matérias do DMA apresentadas, 80% destas possuem índice total de reprovação da turma acima de 50%. Neste contexto a reprovação total da turma significa respectivamente que nenhum aluno matriculado oriundo do Dcomp obteve aprovação na disciplina cursada.

A seguir é possível verificar que a disciplina "Física A" também apresenta índice total de reprovação da turma acima de 50% e somente 12.23% das turmas ofertadas alcançaram a aprovação total dos alunos matriculados. Diante deste cenário o Departamento de Estatística e

Ciências Atuarias - DECAT que oferece a disciplina "Estatística Aplicada" é o que apresenta melhor aproveitamento geral com 30.94% de aprovação total das turmas ofertadas ao Dcomp. Neste cenário a aprovação total da turma significa respectivamente que todos os alunos matriculados nesta matéria e que são oriundos do Dcomp foram aprovados na disciplina cursada.

Neste momento o que se pode inferir é que alunos oriundos do Dcomp possuem déficit em disciplinas exatas. Esta situação pode ser decorrente a diversos fatores, entre elas, a deficiência no ensino básico. Neste contexto é sugerido uma parceria entre Departamentos da Universidade Federal de Sergipe de forma a oferecer mini-cursos voltados as matérias que concentram evasão. Deste modo é esperado recuperar o aluno e evitar a desistência/reprovação na disciplina.

5.4.3 Quais são as principais características de estudantes que largarão o curso selecionado?

Para responder esta questão foram analisados 14.704 registros de alunos matriculados em turmas ofertadas entre 2007 e 2018. Estes dados foram resultantes da análise gerada através do algoritmo árvore de decisão com o valor de acurácia 98,647%. Destes registros, 7.547 são de alunos classificados como reprovados e 7.157 de discentes classificados como aprovados. Dentro deste contexto os atributos observados são o atributo "dificuldade média da turma" (DMT) definido na Seção 5 e o atributo "dificuldade média do aluno" (DMA) definido na Seção 5.3.

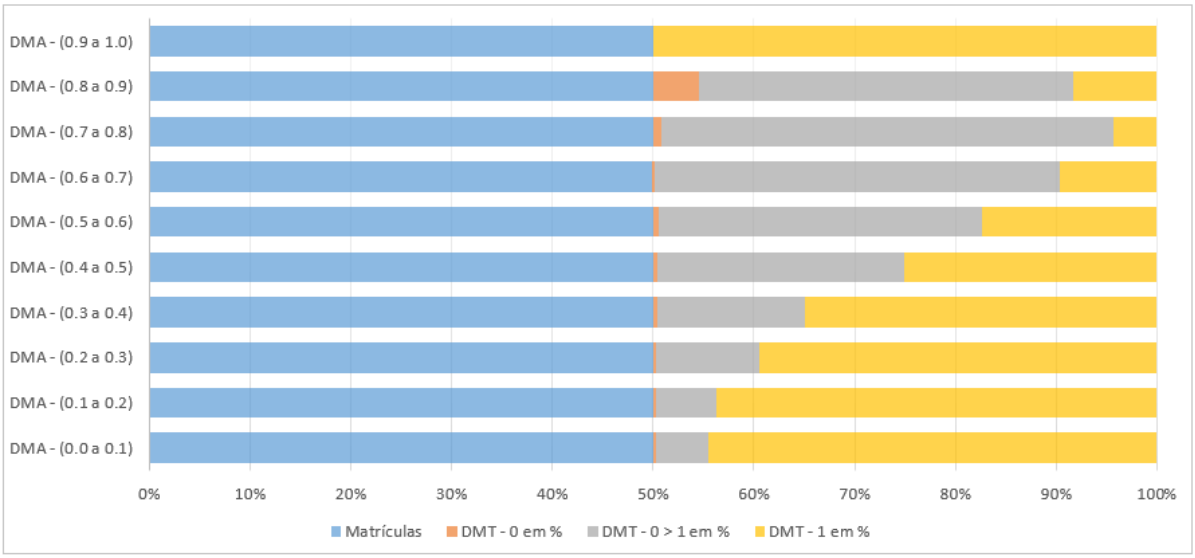
Após a aplicação da técnica de normalização Min-Max [0.0,1.0] os valores assumidos para DMA e DMT respeitam a dimensão estabelecida entre 0 e 1. No contexto de alunos reprovados, o atributo DMT com o valor 0 aponta para registros de alunos onde a nota não é o fator crucial para a reprovação, sendo categorizado como faltas e outras razões características. O atributo DMT com o valor 1 indica que todos os alunos registrados na disciplina foram reprovados. Para as turmas com o valor DMT entre 0 e 1 indica que houveram aprovações e reprovações na disciplina, o que não caracteriza uma discrepância no contexto da matéria ministrada. A Tabela 20 apresenta os valores encontrados no contexto de alunos reprovados.

Tabela 20 – Análise de alunos classificados como reprovados entre 2007 até 2018.

	Matrícula(s)	DMT - 0 em %	DMT - 0 > 1 em %	DMT - 1 em %
DMA - (0.0 a 0.1)	299	0.668	10.367	88.963
DMA - (0.1 a 0.2)	348	0.668	12.068	87.356
DMA - (0.2 a 0.3)	397	0.668	20.654	78.841
DMA - (0.3 a 0.4)	539	0.927	29.128	69.944
DMA - (0.4 a 0.5)	815	0.736	49.202	50.061
DMA - (0.5 a 0.6)	1110	1.081	64.234	34.684
DMA - (0.6 a 0.7)	1700	0.529	80.235	19.235
DMA - (0.7 a 0.8)	2260	1.637	89.601	8.761
DMA - (0.8 a 0.9)	78	8.974	74.358	16.666
DMA - (0.9 a 1.0)	1	0.000	0.000	100.000

A Tabela 20 demonstra que alunos com valor DMA até 0.5 possuem registros em turmas que apresentam mais de 50% de reprovação total de discentes matriculados. Estes alunos não tornam a aparecer sob novo atributo DMA durante este estudo, o que pode indicar evasão em turmas com altos índices de reprovação que já foram apresentadas na Seção 5.4.2. Os alunos que possuem DMA com valor acima de 0.5 estão concentrados em turmas que apresentaram menos reprovações totais de alunos. Estes alunos com DMA acima de 0.5 tornam a aparecer em novas consultas ao SGBD, o que indica, a continuação do curso selecionado pelo aluno apesar da reprovação em disciplina matriculada. A Figura 31 apresenta a distribuição de alunos reprovados por turmas DMT 1 e turmas com DMT entre 0 e 1.

Figura 31 – Distribuição de alunos reprovados em turmas DMT 1 e em turmas DMT entre 0 e 1.



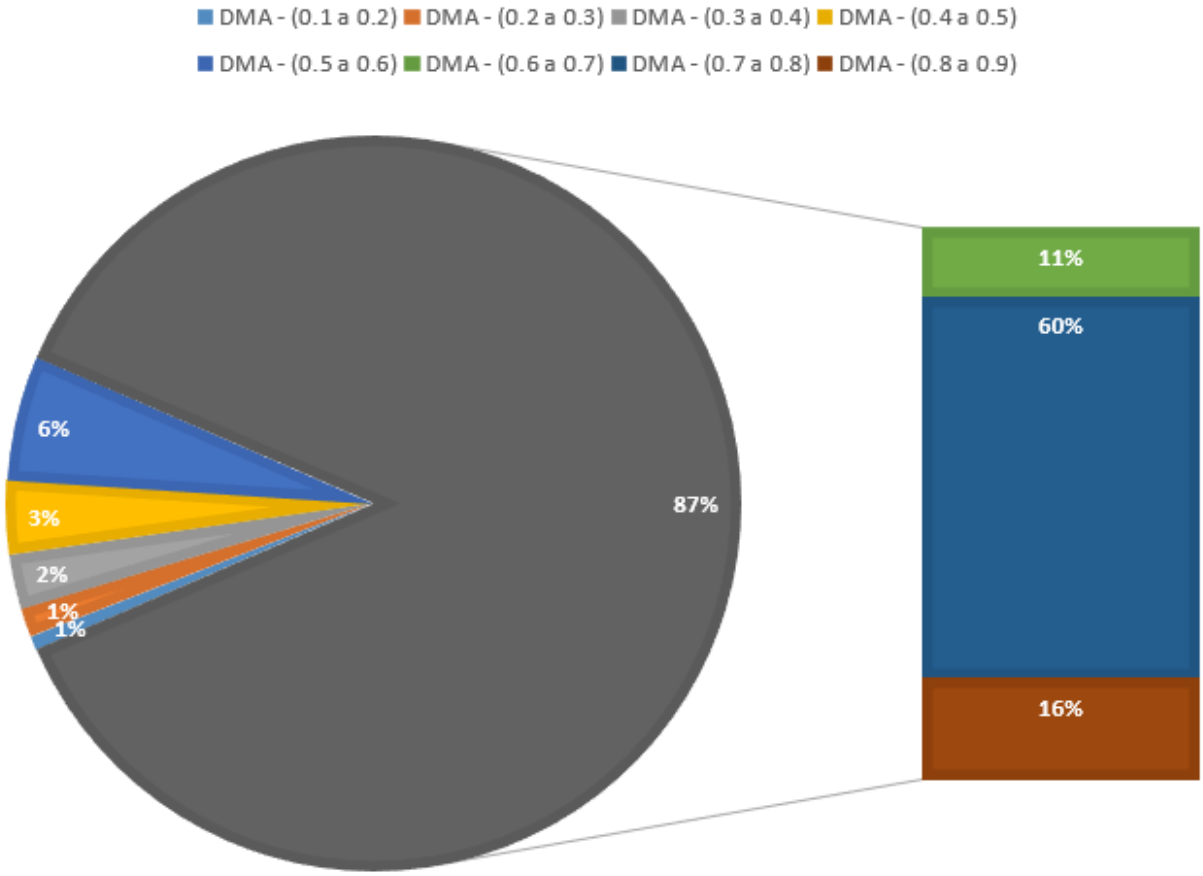
No contexto de alunos aprovados, o atributo DMT com o valor 0 indica que houve aprovação total da turma, ou seja, todos os discentes registrados foram aprovados. O valor DMT entre 0 e 1 indica que houve aprovações e reprovações na turma. A Tabela 21 apresenta os valores encontrados quando observado o contexto dos alunos aprovados.

Tabela 21 – Análise de alunos classificados como aprovados entre 2007 até 2018.

	Matrícula(s)	DMT - 0 em %	DMT - 0 > 1 em %
DMA - (0.0 a 0.1)	1	100.000	0.000
DMA - (0.1 a 0.2)	44	18.181	81.818
DMA - (0.2 a 0.3)	91	30.769	69.230
DMA - (0.3 a 0.4)	170	30.588	69.411
DMA - (0.4 a 0.5)	232	29.741	70.258
DMA - (0.5 a 0.6)	391	24.552	75.447
DMA - (0.6 a 0.7)	773	16.688	83.311
DMA - (0.7 a 0.8)	4319	20.652	79.347
DMA - (0.8 a 0.9)	1133	20.035	79.964
DMA - (0.9 a 1.0)	1	100.000	0.000

A Tabela 21 apresenta que alunos com índice DMA entre 0.6 a 0.9 apresentam concentração de matrículas e aprovações nas turmas selecionadas. Apesar de alunos com DMA menor a 0.6 apresentarem aprovações em turmas, o índice DMA destes representam 13% do conjunto de alunos aprovados. Esta informação pode indicar que o discente possui aprovações porém ainda possui reprovações suficientes para impactar o atributo DMT. Caso o discente esteja matriculado em disciplinas com altos índices de reprovação, como apresentado na Seção 5.4.2, pode ocorrer a desistência do curso selecionado impulsionando ainda mais a evasão. A Figura 32 apresenta distribuição de alunos aprovados por DMA.

Figura 32 – Alunos aprovados entre 2007 a 2018 distribuídos por DMA.



Os subgrupos segregados através do atributo DMA demonstram que existe um crescimento gradativo na quantidade de alunos que se afastam do valor DMA 0. O que leva a assumir que quanto maior o índice DMA menor é a possibilidade da evasão deste estudante.

Neste momento o que se pode inferir é que alunos que possuem valor DMA alto evadem menos do que alunos que possuem DMA baixo. Isto significa que alunos que sempre obtém aprovações permanecerão no curso selecionado apesar de obter reprovações em determinado momento de vida acadêmica. Já alunos que possuem valor DMA baixo em turmas com indíce de dificuldade maior possuem alta propensão a evadir do curso caso não exista nenhum programa

da Universidade a recuperá-los.

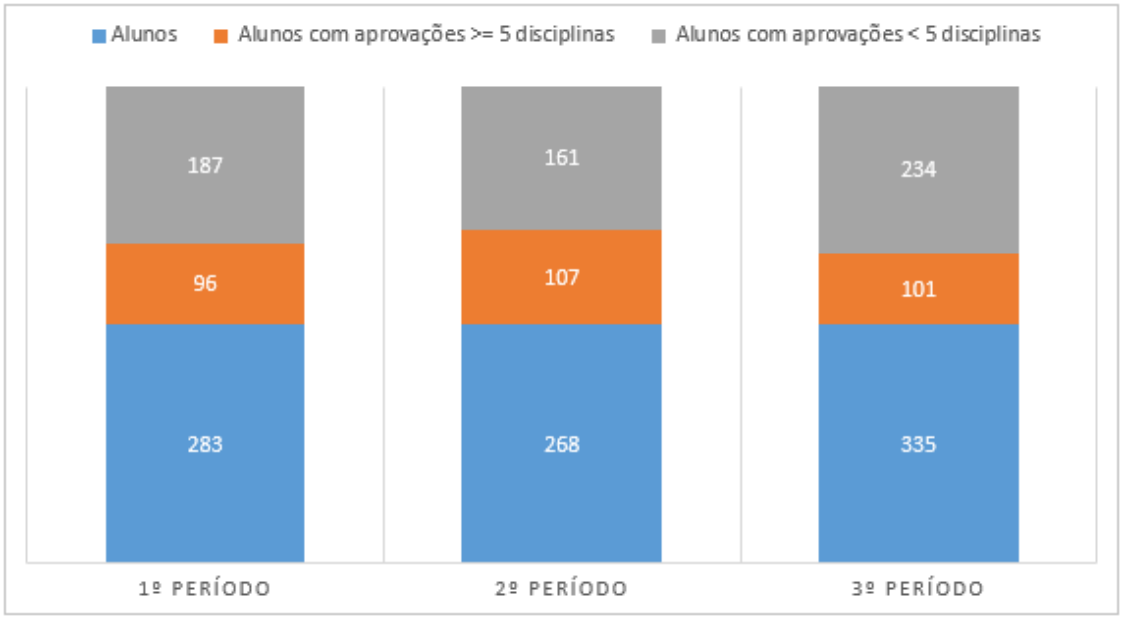
5.4.4 A quantidade de matérias selecionadas influencia o desempenho do aluno e do algoritmo que está sendo aplicado?

Para responder a esta questão de pesquisa foram analisadas 11.126 aprovações de alunos matriculados em turmas de 1º a 3º períodos dos cursos de Sistemas de Informação, Ciência da Computação e Engenharia da Computação ofertadas entre 2007 e 2018. Em seguida foi contabilizado a quantidade de disciplinas de 1º ao 3º período obrigatórios de cada curso. A partir deste levantamento detectou-se que a grade de cada curso ofertado pelo Dcomp apresenta de 5 a 8 componentes curriculares obrigatórios por período. Deste modo foram isolados cada curso, seguimentado através do período, e observados as aprovações de cada subconjunto de acordo com os componentes curriculares obrigatórios. A Tabela 22 e a Figura 33 demonstram as aprovações e quantidade de alunos referentes ao curso de Sistemas de Informação - SI respectivamente.

Tabela 22 – Aprovações em turmas de 1º a 3º período de SI entre 2007 até 2018.

Período	Aprovações >= 5 Matérias em %	Aprovações < 5 Matérias em %	Aprovações
1	33.92%	66.07%	1024
2	39.92%	60.07%	989
3	30.14%	69.85%	1035

Figura 33 – Alunos aprovados em turmas de 1º a 3º período de SI entre 2007 até 2018.



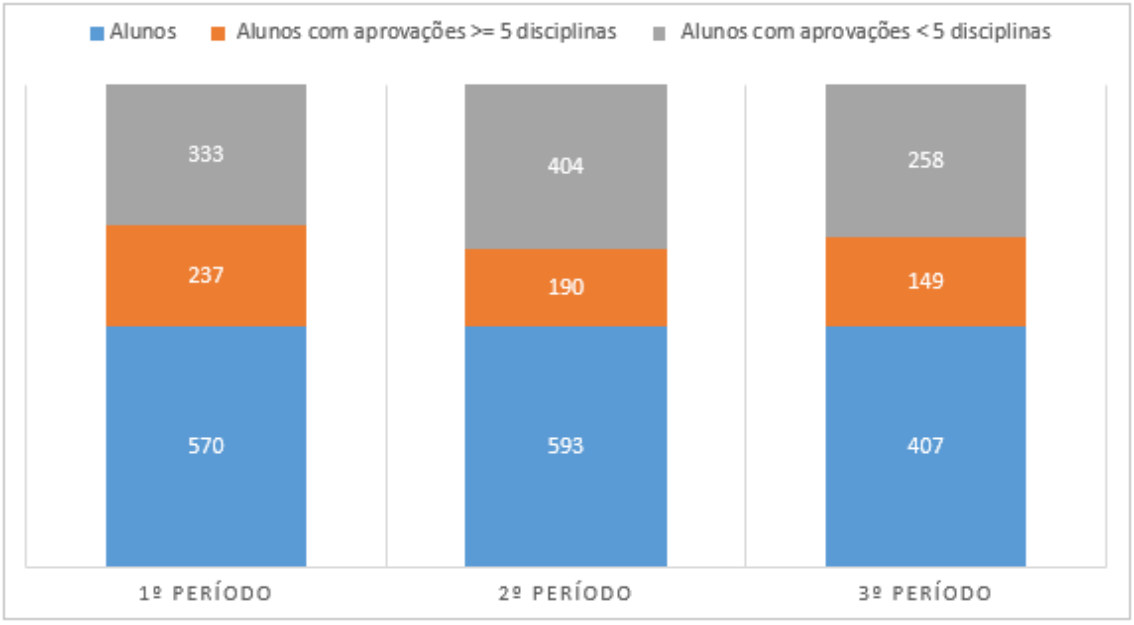
No contexto da grade de SI foram analisados 886 alunos equivalentes a 3.048 aprovações entre 1º a 3º período. Destes discentes, 304 conseguiram aproveitamento maior ou igual a 5 disciplinas de acordo com a grade curricular instituída pelo curso. Estes alunos representam 34.31% dos discentes analisados. Em contrapartida, 582 estudantes não obtiveram a aprovação

esperada correspondente a sua grade curricular, ou seja, os alunos apresentaram perda de uma ou mais disciplinas obrigatórias instituídas no período cursado. Os estudantes que possuem aproveitamento menor a 5 disciplinas por período representam 65.68% dos alunos analisados no curso de SI. A seguir, a Tabela 23 e a Figura 34 demonstram as aprovações e quantidade de alunos referentes ao curso de Ciência da Computação - CC respectivamente.

Tabela 23 – Aprovações em turmas de 1º a 3º período de CC entre 2007 até 2018.

Período	Aprovações >= 5 Matérias em %	Aprovações < 5 Matérias em %	Aprovações
1	41.57%	58.42%	2104
2	32.04%	68.12%	1977
3	36.60%	63.39%	1466

Figura 34 – Alunos aprovados em turmas de 1º a 3º período de CC entre 2007 até 2018.

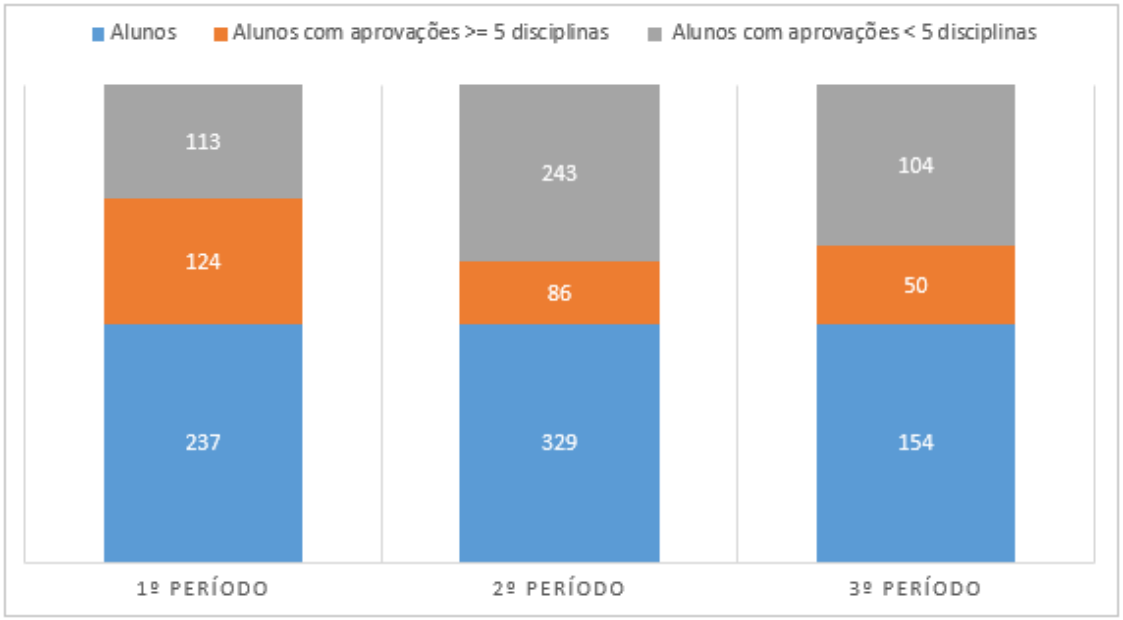


No contexto da grade de CC foram analisados 1570 alunos equivalentes a 5.547 aprovações entre 1º a 3º período. Destes discentes, 576 conseguiram aproveitamento maior ou igual a 5 disciplinas, o que representa 36.68% dos alunos. Em contrapartida, 995 estudantes não obtiveram a aprovação esperada correspondente a sua grade curricular, o que representa 63.37% dos alunos analisados do curso de CC. Em seguida a Tabela 24 e a Figura 35 apresentam as aprovações e quantidade de alunos referentes ao curso de Engenharia da Computação - EC respectivamente.

Tabela 24 – Aprovações em turmas de 1º a 3º período de EC entre 2007 até 2018.

Período	Aprovações >= 5 Matérias em %	Aprovações < 5 Matérias em %	Aprovações
1	52.32%	47.67%	969
2	26.13%	73.86%	1067
3	32.46%	67.53%	495

Figura 35 – Alunos aprovados em turmas de 1º a 3º período de EC entre 2007 até 2018.



No contexto da grade de EC foram analisados 720 alunos equivalentes a 2.531 aprovações entre 1º a 3º período. Destes discentes, 260 conseguiram aproveitamento maior ou igual a 5 disciplinas de acordo com a grade curricular instituída pelo curso o que corresponde a 36.11% dos discentes analisados. Em contrapartida, 460 estudantes não obtiveram a aprovação esperada correspondente a sua grade curricular. Os estudantes que possuem aproveitamento menor a 5 disciplinas por período representam 63.88% dos alunos analisados do curso EC.

Ao observar os 3 cursos do Dcomp constatou-se que dos 3.176 alunos selecionados apenas 35.89% apresentam rendimento maior a 5 disciplinas por período obrigatório, o que corresponde a 1.140 estudantes. A perda de disciplinas no início do curso é uma das possíveis causas que impulsionam a evasão conforme a Seção 5.4.4. Deste modo 64.10% dos discentes, o que representa 2.036 alunos, estariam suscetíveis a evasão gradual do curso selecionado.

Neste momento o que se pode inferir é que a quantidade de matérias influencia o desempenho dos alunos visto que, nos cursos e períodos observados a aprovação maior a cinco matérias referente a grade obrigatória se manteve menor a 50%. Nesta situação as ações sugeridas na Seção 5.4.2 podem auxiliar parcialmente o desempenho dos discentes além de incentivar a permanência na disciplina matriculada. Dentro deste estudo o algoritmo árvore de decisão não apresentou variantes para classificação, o que ressalta a importância da criação do atributo "dificuldade média do aluno" para representar esta informação e fomentar a acurária do algoritmo.

5.4.5 A correlação de matérias por semestre influencia o desempenho do aluno?

Para responder a esta questão de pesquisa foram analisadas as matérias obrigatórias de 1º a 3º período dos cursos de Ciência da Computação, Sistemas de Informação e Engenharia

da Computação de grades ofertadas em 2007, 2008 e 2009 respectivamente. Este levantamento considerou os valores do atributo "dificuldade média da turma" para representar a dificuldade das disciplinas obrigatórias por período a ser cursado pelo aluno.

Para esta análise foram contabilizadas as turmas ofertadas entre 2007 e 2018 de acordo as grades correspondentes aos cursos do Dcomp. Posteriormente foi calculada a média DMT de todas as turmas por disciplina considerando o atributo DMT individual de cada uma destas. Nesta etapa não foi aplicada a técnica de normalização Min-Max [0.0,1.0], deste modo, os valores dos atributos DMTs seguem o contexto de variação do número 1 ao 30. O número 1 indica que todos os alunos matriculados na turma foram aprovados, e em contraparte, o número 30 indica que todos os alunos registrados na disciplina foram reprovados.

Esta representação do valor médio DMT para cada matéria visa atribuir um valor de dificuldade correspondente ao histórico de cada disciplina, assim, é possível avaliar a correlação das matérias por período e por curso. A seguir, a Tabela 25 apresenta o levantamento correspondente ao curso de Ciência da Computação em relação a grade curricular ofertada em 2007.

Tabela 25 – 1º a 3º período de CC da grade curricular ofertada em 2007.

Período	Disciplina	Valor DMT
1º	FUNDAMENTOS DA COMPUTAÇÃO	6.01
1º	PROGRAMAÇÃO IMPERATIVA	15.51
1º	MÉTODOS E TÉCNICAS DE PESQUISA	9.07
1º	CÁLCULO I	21.68
1º	VETORES E GEOMETRIA ANALÍTICA	21.38
1º	FUNDAMENTOS DE MATEMATICA PARA COMPUTACAO	11.91
2º	PROGRAMAÇÃO ORIENTADA A OBJETOS	12.12
2º	ESTRUTURA DE DADOS I	11.75
2º	CIRCUITOS DIGITAIS I	4.00
2º	LABORATÓRIO DE CIRCUITOS DIGITAIS I	4.00
2º	LÓGICA PARA COMPUTAÇÃO	6.91
2º	FÍSICA A	17.94
2º	CÁLCULO II	17.76
3º	PROGRAMAÇÃO DECLARATIVA	4.72
3º	PROGRAMAÇÃO PARA WEB	10.20
3º	ESTRUTURA DE DADOS II	5.25
3º	ARQUITETURA DE COMPUTADORES I	9.97
3º	INFORMÁTICA, ÉTICA E SOCIEDADE	2.13
3º	FÍSICA B	12.69
3º	ÁLGEBRA LINEAR I	17.58

De acordo com a Tabela 25 é possível verificar que as disciplinas apresentam valores DMT distintos entre si. Neste contexto as matérias que possuem muitas reprovações e histórico condizente a este resultado se destacam através de altos valores DMT. Com esta informação é possível calcular a dificuldade média do período - DMP.

A DMP pode ser representada através da média dos valores DMT das disciplinas obrigatórias a serem cursadas no período. É uma representação criada para auxiliar o entendimento do cenário formado pelas disciplinas que compõem o período obrigatório curricular. A Tabela 26 apresenta os valores DMP correspondentes ao curso de CC entre o 1º e 3º período.

Tabela 26 – DMP do 1º ao 3º período de de CC da grade curricular ofertada em 2007.

Curso	DMP - 1º Período	DMP - 2º Período	DMP - 3º Período
CC	14.26	10.64	8.93

A Tabela 26 reflete que o 1º período de CC possui maior complexidade para o aluno em relação a outros períodos que apresentam valor DMP decrescente. Esta redução é consequência direta da correlação das disciplinas que compõe a grade curricular obrigatória do curso de CC. A seguir a Tabela 27 apresenta o levantamento correspondente ao curso de SI em relação a grade curricular ofertada em 2008 e a Tabela 28 demonstra os valores DMP encontrados.

Tabela 27 – Turmas de 1º a 3º período de SI ofertada em 2008.

Período	Disciplina	Valor DMT
1º	INTRODUÇÃO A ADMINISTRAÇÃO	10.55
1º	PROGRAMAÇÃO IMPERATIVA	15.51
1º	FUNDAMENTOS DA COMPUTAÇÃO PARA SISTEMAS DE INFORMAÇÃO	9.50
1º	CÁLCULO I	21.68
1º	FUNDAMENTOS DE ÁLGEBRA PARA COMPUTAÇÃO	6.31
2º	ORGANIZAÇÃO, MÉTODOS E SISTEMAS ADMINISTRATIVOS	5.35
2º	PROGRAMAÇÃO ORIENTADA A OBJETOS	12.12
2º	ORGANIZAÇÃO E ARQUITETURA DE COMPUTADORES	5.51
2º	INGLÊS INSTRUMENTAL	7.50
2º	VETORES E GEOMETRIA ANALÍTICA	21.38
2º	FUNDAMENTOS DE MATEMÁTICA	14.01
3º	SOCIOLOGIA DAS ORGANIZAÇÕES	7.00
3º	ESTRUTURA DE DADOS PARA SISTEMAS DE INFORMAÇÃO I	7.34
3º	TEORIA DA COMPUTAÇÃO	7.04
3º	TEORIA GERAL DOS SISTEMAS	1.93
3º	ESTATÍSTICA APLICADA	11.86
3º	INTRODUÇÃO A METODOLOGIA CIENTÍFICA	14.85

Tabela 28 – DMP do 1º ao 3º período de SI da grade curricular ofertada em 2008.

Curso	DMP - 1º Período	DMP - 2º Período	DMP - 3º Período
SI	12.71	10.97	8.33

A Tabela 27 apresenta valores DMT menores em relação ao curso de CC. A grade de SI possui disciplinas mais diversificadas com menores índices de reprovação. A Tabela 28 reflete que o 1º período de SI possui maior complexidade para o aluno em relação a outros períodos de valor DMP decrescente. Esta redução do DMP é consequência direta da correlação das disciplinas que compõe a grade curricular obrigatória do curso de SI. A seguir a Tabela 29 apresenta o levantamentodo do curso de EC em relação a grade curricular ofertada em 2009 e a Tabela 30 demonstra os valores DMP correspondentes.

Tabela 29 – Turmas de 1º a 3º período de EC ofertada em 2009.

Período	Disciplina	Valor DMT
1º	FUNDAMENTOS DE ENGENHARIA DE COMPUTAÇÃO	4.72
1º	PROGRAMAÇÃO IMPERATIVA	15.51
1º	DESENHO TÉCNICO	17.93
1º	CÁLCULO I	21.68
1º	VETORES E GEOMETRIA ANALÍTICA	21.38
1º	FUNDAMENTOS DE MATEMÁTICA PARA COMPUTAÇÃO	11.91
2º	PROGRAMAÇÃO ORIENTADA A OBJETOS	12.12
2º	METODOLOGIA E COMUNICAÇÃO CIENTÍFICA	15.60
2º	RESISTÊNCIA DOS MATERIAIS	19.79
2º	FÍSICA A	17.94
2º	LABORATÓRIO DE FÍSICA A	13.16
2º	INGLÊS INSTRUMENTAL	7.50
2º	CÁLCULO II	17.76
2º	ÁLGEBRA LINEAR I	17.58
3º	ESTRUTURA DE DADOS PARA ENGENHARIA DE COMPUTAÇÃO	5.94
3º	CIRCUITOS DIGITAIS	18.04
3º	FÍSICA B	12.69
3º	LABORATÓRIO DE FÍSICA B	7.92
3º	CÁLCULO III	17.74
3º	EQUAÇÕES DIFERENCIAIS ORDINÁRIAS	11.14

Tabela 30 – DMP do 1º ao 3º período de EC da grade curricular ofertada em 2009.

Curso	DMP - 1º Período	DMP - 2º Período	DMP - 3º Período
EC	15.52	15.18	12.24

A Tabela 29 apresenta valores DMT maiores em relação ao curso de CC e SI. A grade de EC possui muitas disciplinas iniciais ofertadas pelo Departamento de Matemática - DMAI e Departamento de Física - DFI, ambos com altos valores DMT, o que eleva os resultados DMP correlacionados. A Tabela 30 reflete que o 1º período de EC possui maior complexidade para o

aluno em relação a outros períodos de valor DMP decrescente. A redução DMP possui relação direta com as disciplinas que compõe a grade curricular obrigatória do curso de EC.

Ao analisar a grade dos cursos ofertados pelo Dcomp e os DMTs correspondentes ao 1º e 3º período destes é possível identificar que o 1º período de todos os cursos apresenta maior valor DMP em relação aos outros. Este valor pode ter associação direta com as informações refletidas na Seção 5.4.3. Os altos valores DMP retratam diretamente a correlação das disciplinas da grade curricular obrigatória. Algumas disciplinas indicadas na Seção 5.4.2 reaparecem nesta análise por fazer parte do conjunto de matérias com maior índice de reprovação e que consequentemente impulsionam a evasão. É importante salientar que as disciplinas apresentadas na Seção 5.4.2 se concentram entre o 1º e 3º período dos cursos ofertados pelo Dcomp.

Neste momento o que se pode inferir é que para o aluno os três primeiros períodos do curso são os mais exigentes, devido ao grau de dificuldade do conjunto de matérias que compõem a grade curricular. Apesar dos períodos apresentarem atenuação de dificuldade no decorrer do curso a dificuldade do conjunto de disciplinas continua alta, o que pode fomentar a evasão.

5.4.6 Quais são os motivos que impulsionam a evasão dos alunos de cursos de computação?

Para responder a esta questão de pesquisa foram coletadas informações segmentadas das Seções 5.4.2 a 5.4.5. No total foram avaliados 4.017 alunos e 49.013 registros de matrículas em turmas disponíveis diversificados entre os cursos de SI, CC e EC do Dcomp. Esta análise considerou os atributos "dificuldade média da turma", "dificuldade média do aluno" e "dificuldade média do período" para representar numericamente estas grandezas.

Na Seção 5.4.2 é destacado que as disciplinas que mais reprovam é composta por 70% de componentes curriculares que não são ofertados pelo Departamento de Computação. Esta grandeza pode indicar alta deficiência dos estudantes em relação à matemática, física e outras disciplinas correlacionadas. Uma possibilidade para mitigar a evasão destas matérias é a criação de projetos no Dcomp com atividades paralelas que reforcem o conteúdo deficiente dos alunos.

Na Seção 5.4.3 é citado que alunos que possuem DMA com valor acima de 0.5 aparecem em muitas turmas demonstrando a continuidade no curso selecionado, em contrapartida, discentes que apresentam valor DMA menor a 0.5 gradativamente deixam de aparecer no SGBD sugerindo a evasão do curso selecionado. Existe alta concentração de alunos com DMA abaixo de 0.5 em disciplinas que apresentam reprovação total da turma, o que pode ser um indicador, para alunos matriculados em disciplinas que possuem maior índice de reprovação conforme Seção 5.4.2. Com este resultado alunos que possuem valor DMA alto evadem menos do que alunos que possuem DMA baixo. Isto significa que alunos que sempre obtém aprovações permanecerão no curso selecionado apesar de obter reprovações em determinado momento de vida acadêmica. Já alunos que possuem valor DMA baixo em turmas com índice de dificuldade maior possuem

alta propensão a evadir do curso caso não exista nenhum programa de recuperação destes na Universidade.

Na Seção 5.4.4 foram analisados 11.126 registros de aprovações de alunos matriculados em turmas de 1º a 3º períodos em cursos do Dcomp entre os anos 2007 a 2018. Neste contexto, 35.89% dos alunos apresentaram rendimento maior a 5 disciplinas cursadas em período obrigatório. Esta informação retrata que 64.10% dos discentes apresentaram perda em uma ou mais disciplinas de seu período curricular correspondente. Estes números indicam que a quantidade de matérias cursadas em paralelo pode influenciar o desempenho do aluno e consequentemente fomentar a evasão gradual do curso. Nesta situação as ações sugeridas na Seção 5.4.2 podem auxiliar parcialmente o desempenho dos discentes além de incentivar a permanência na disciplina matriculada.

Na Seção 5.4.5 é possível mensurar a DMP através da média das DMTs disponíveis por período. Os altos valores DMP apresentados estão diretamente correlacionados ao conjunto de disciplinas que formam a grade curricular obrigatória. A tendência é que a DMP se torne mais amena a medida que o discente avança no curso e obtém mais aprovações. É importante ressaltar que o 1º período para todos os cursos é o que apresenta maior dificuldade para o aluno e que as disciplinas apresentadas na Seção 5.4.2 se concentram entre o 1º e 3º período dos cursos ofertados pelo Dcomp. Para o aluno, os três primeiros períodos do curso são os mais exigentes devido ao grau de dificuldade do conjunto de matérias que compõem a grade curricular. Apesar dos períodos apresentarem atenuação da dificuldade no decorrer do curso a dificuldade do conjunto de disciplinas obrigatórias continua alta, o que pode fomentar a evasão.

6

Conclusões e Trabalhos Futuros

Nesta pesquisa explorou-se a transformação de dados, que estavam a priori armazenados em CSVs acadêmicos, em informações potencialmente úteis para apoiar a mitigação da evasão através da identificação de perfis evasivos com o auxílio de técnicas de mineração de dados.

Conforme verificado neste trabalho, a evasão ainda é um fenômeno em crescimento no Dcomp, bem como na UFS como um todo, justificando a necessidade de identificar padrões evasivos, analisá-los e posteriormente elaborar planos para incentivar a permanência do aluno no curso selecionado.

Este trabalho apresentou uma abordagem para selecionar os melhores atributos para classificação, aplicado na predição da evasão escolar focada na aprovação ou reprovação da disciplina matriculada, utilizando a criação de atributos. Esta abordagem contribuiu para o processo de identificação de padrões a serem utilizados na predição da evasão dos cursos da UFS, através de um estudo de caso.

A arquitetura de *Datawarehouse* Pentaho (NETO, 2016) implementada em ambiente da UFS auxiliou as tarefas de limpeza, extração, transformação e carga dos dados, passos importantes na tarefa de mineração de dados, além de permitir a geração de relatórios analíticos dos dados através dos *Reports* criados sobre a evasão, disponibilizados aos gestores educacionais do Dcomp.

Entre as técnicas de mineração de dados educacionais estudadas as melhores acurácias foram obtidas através do algoritmo árvore de decisão. Na escolha dos melhores atributos para a tarefa de mineração, foram criados os atributos “dificuldade média do aluno” e “dificuldade média da turma” para otimizar a acurácia do algoritmo de classificação, tendo como diferencial a agregação um componente coletivo ao desempenho individual do aluno e da disciplina a ser cursada.

Na análise da evasão, foi constatado que há fortes indicativos que fomentam a evasão até

o 3º período do curso, independente do total de períodos que seja composto a grade curricular. Isto reduz o escopo a ser analisado e permite concentrar os esforços em medidas para mitigar a evasão de alunos com características identificadas em perfis evasivos.

A abordagem computacional proposta nesta pesquisa possui vantagens, como a de se gerar inferências sobre o conjunto de alunos matriculados, mas possui também desvantagens, por apresentar os sintomas que podem ocultar as reais causas da evasão. Uma possibilidade para obter resultados através da mineração de dados pode ser a coleta de outras informações dos alunos, através de questionários específicos, uso de ambientes virtuais de aprendizagem, entre outros, tendo o cuidado para armazenar essas informações ao longo do tempo.

Com a abordagem computacional proposta para a identificação de perfis evasivos, permite-se visualizar padrões que podem auxiliar a tomada de decisão de professores e gestores educacionais, além de avaliar a possibilidade da evasão de cada aluno por disciplina matriculada, abrangendo desde o nível estratégico até o nível operacional com o uso de técnicas de mineração de dados.

6.1 Limitações

As questões elaboradas para este estudo podem não compreender todos os pontos que compõem o tema *Education Data Mining* voltados à evasão no ensino superior. As construções destas perguntas foram baseadas diretamente nas dores do DCOMP, no intuito de fomentar o interesse à pesquisa e a evidenciar oportunidades para novos modelos de investigação ainda pouco explorados.

Outra limitação desta pesquisa abrange a criação dos novos atributos propostos na abordagem computacional. Apesar destes novos atributos recém-criados possibilitarem uma acurácia otimizada do algoritmo árvore de decisão, a manipulação e interpretação desta informação demanda conhecimento focado em técnicas de análise de dados. O que pode restringir a replicação desta abordagem para pesquisadores sem a devida especialização.

Além deste pontos, é importante ressaltar que esta pesquisa alcançou uma quantidade restrita de bases durante a realização do mapeamento sistemático, ou seja, estudos significativos podem não ter sido contemplados. Para complementar esta visão o período estipulado pode não abranger importantes trabalhos sobre o assunto.

6.2 Trabalhos Futuros

Analisar a evasão apenas utilizando os resultados alcançados através da mineração de dados educacionais pode ser insuficiente sob o ponto de vista gerencial. Entende-se como importante unir os resultados de mineração de dados com análises multidimensionais em um *Datawarehouse* (NETO, 2016). Dessa forma pretende-se trabalhar na inclusão de mais *Reports*

no *Datawarehouse* e a criação *Dashboards* com perfis para diferentes níveis de gestores educacionais, criando alertas para possibilitar a intervenção junto ao aluno em tempo oportuno para mitigar a evasão do curso selecionado.

No contexto da mineração de dados educacionais a pesquisa também pode ser ampliada. Pretende-se aplicar a abordagem proposta em outras bases de dados, bem como, realizar o cruzamento de informações do Exame Nacional do Ensino Médio - ENEM em relação ao 1º período dos cursos do Dcomp. Deste modo já é possível ter uma medida voltada a alunos no início do curso e evitar a possibilidade de evasão.

É importante ampliar o escopo para tratar a evasão em conjunto com o problema gerado pela retenção de alunos com perdas em disciplinas de componente curricular obrigatório. Esta retenção faz com que os graduandos demorem tempo superior ao previsto a grade curricular definida, o que pode consequentemente, desistimular o aluno a concluir o curso selecionado.

Sugere-se realizar um projeto multidisciplinar com o Departamento de Educação - DED e o Departamento de Psicologia - DPS da Universidade Federal de Sergipe, para aprofundamento deste estudo, com o intuito de minimizar as deficiências relacionadas a uma abordagem apenas computacional.

Neste aspecto, este trabalho pode ser considerado como uma análise inicial, que pode ser utilizada para fornecer dados quantitativos para uma exploração mais ampliada sob o ponto de vista do problema da evasão em cursos superiores de computação da Universidade Federal de Sergipe.

Referências

A. FACELI K., L. A. G. J. C. Inteligência artificial – uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, 2011. Citado na página 51.

A. TOLOSI L., S. O. A.; T., L. Permutation importance: a corrected feature importance measure. *Bioinformatics*, Oxford University Press, v. 26, n. 10, p. 1340–1347, 2010. Citado na página 51.

ABREU, F. S. G. G. Desmistificando o conceito de etl. *Revista de Sistemas de Informação*, 2008. Citado na página 17.

ADEODATO, P. et al. Neural networks vs logistic regression: a comparative study on a large data set. In: . [S.l.: s.n.], 2004. p. 355–358. Citado na página 39.

AL-SHABANDAR, R. et al. Machine learning approaches to predict learning outcomes in massive open online courses. In: IEEE. *2017 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2017. p. 713–720. Citado 5 vezes nas páginas 11, 34, 35, 36 e 37.

ALGARNI, A. Data mining in education. *International Journal of Advanced Computer Science and Applications*, v. 7, 06 2016. Citado 2 vezes nas páginas 23 e 24.

AMBIEL, R. A. M. Construção da Escala de Motivos para Evasão do Ensino Superior. *Avaliação Psicológica*, scielopepsic, v. 14, p. 41 – 52, 04 2015. ISSN 1677-0471. Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712015000100006&nrm=iso>. Citado na página 10.

AMERI, S. et al. Survival analysis based framework for early prediction of student dropouts. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. [S.l.: s.n.], 2016. p. 903–912. Citado 3 vezes nas páginas 33, 34 e 36.

ANALYTICS, C. *Anaconda Software Distribution, v. 2.4. 0*. 2016. Citado na página 51.

ANGRA, S.; AHUJA, S. Implementation of data mining algorithms on student's data using rapid miner. In: IEEE. *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. [S.l.], 2017. p. 387–391. Citado 2 vezes nas páginas 34 e 36.

ANTUNES, C. *Anticipating student's failure as soon as possible*. [S.l.]: CRC Press, 2010. v. 353. Citado na página 40.

ANTUNES, R. et al. Análise de integração de mineradores de dados com a plataforma interimage – qual a melhor solução? *Revista Brasileira de Cartografia*, v. 70, p. 1470–1509, 12 2018. Citado na página 30.

ATHANI, S. S. et al. Student performance predictor using multiclass support vector classification algorithm. In: IEEE. *2017 International Conference on Signal Processing and Communication (ICSPC)*. [S.l.], 2017. p. 341–346. Citado 3 vezes nas páginas 34, 36 e 37.

AZUAJE, F. *Witten ih, frank e: Data mining: Practical machine learning tools and techniques 2nd edition*. [S.l.]: BioMed Central, 2006. Citado na página 61.

BAKER, R. S.; YACEF, K. The state of educational data mining in 2009: A review and future visions. *JEDM| Journal of Educational Data Mining*, v. 1, n. 1, p. 3–17, 2009. Citado na página 23.

BALANIUK, R. et al. Predicting evasion candidates in higher education institutions. In: SPRINGER. *International Conference on Model and Data Engineering*. [S.l.], 2011. p. 143–151. Citado 4 vezes nas páginas 11, 34, 36 e 37.

BASTOS, D. G.; NASCIMENTO, P. S.; LAURETTO, M. S. Proposta e análise de desempenho de dois métodos de seleçao de caracterısticas para random forests. *IX Simpósio Brasileiro de Sistemas de Informaç ao*, p. 49–60, 2013. Citado na página 21.

BEISKEN, S. et al. Knime-cdk: Workflow-driven cheminformatics. *BMC bioinformatics*, Springer, v. 14, n. 1, p. 257, 2013. Citado na página 29.

BENABLO, C. I. P. et al. Higher education student's academic performance analysis through predictive analytics. In: *Proceedings of the 2018 7th International Conference on Software and Computer Applications*. [S.l.: s.n.], 2018. p. 238–242. Citado 3 vezes nas páginas 34, 36 e 37.

BERTHOLD, M. et al. Knime: the konstanz information miner. *First publ. in: Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 79, 2007. New York: Springer, 2008*, V, 11 2009. Citado na página 29.

BOGARÍN, A. et al. Clustering for improving educational process mining. In: *Proceedings of the fourth international conference on learning analytics and knowledge*. [S.l.: s.n.], 2014. p. 11–15. Citado 4 vezes nas páginas 11, 34, 36 e 37.

BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, v. 30, n. 7, p. 1145–1159, 1997. Citado na página 26.

BRAJKOVIĆ, E.; RAKIĆ, K.; KRALJEVIĆ, G. Application of data mining in e-leaming systems. In: IEEE. *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*. [S.l.], 2018. p. 1–5. Citado 3 vezes nas páginas 34, 36 e 37.

CALDERS, T.; PECHENIZKIY, M. Introduction to the special section on educational data mining. *Acm Sigkdd Explorations Newsletter*, ACM New York, NY, USA, v. 13, n. 2, p. 3–6, 2012. Citado 2 vezes nas páginas 34 e 36.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, p. 1–29, 2009. Citado 2 vezes nas páginas 22 e 23.

CEDERBERG, N. *Strategic Alignment of Business Intelligence – A Case Study*. Dissertação (Mestrado) — School of Management, 2010. Citado 2 vezes nas páginas 18 e 19.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado 2 vezes nas páginas 28 e 61.

CHEN, H.-c.; CHIANG, R.; STOREY, V. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, v. 36, p. 1165–1188, 12 2012. Citado na página 19.

COSTA, E. et al. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, v. 1, n. 1, p. 1–29, 2013. Citado na página 20.

de O. Santos, K. J. et al. Supervised learning in the context of educational data mining to avoid university students dropout. In: *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.: s.n.], 2019. v. 2161-377X, p. 207–208. Citado na página 50.

DEEPAK, E. et al. Svm kernel based predictive analytics on faculty performance evaluation. In: IEEE. *2016 International Conference on Inventive Computation Technologies (ICICT)*. [S.l.], 2016. v. 3, p. 1–4. Citado 3 vezes nas páginas 34, 36 e 37.

DEVLIN, B. Data warehouse : from architecture to implementation / b. devlin. 01 1997. Citado na página 18.

DIGIAMPIETRI, L.; NAKANO, F.; LAURETTO, M. Mineração de dados para identificação de alunos com alto risco de evasão: Um estudo de caso. *Grad+ Revista de Graduação da USP*, v. 1, p. 17–23, 07 2016. Citado na página 39.

DWIVEDI, S.; ROSHNI, V. K. Recommender system for big data in education. In: IEEE. *2017 5th National Conference on E-Learning & E-Learning Technologies (ELELTECH)*. [S.l.], 2017. p. 1–4. Citado 2 vezes nas páginas 34 e 36.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996. Citado 3 vezes nas páginas 25, 60 e 62.

FILHO, R. L. L. S. e. a. A evasão no ensino superior brasileiro. *Cadernos de pesquisa, SciELO Brasil v. 37, n. 132*, p. 641 – 659, 2007. Citado 4 vezes nas páginas 10, 11, 13 e 20.

GARTNER. Business intelligence (bi), disponível em: <http://www.gartner.com/it-glossary/business-intelligence-bi/>. 2018. Citado na página 17.

GOPALAKRISHNAN, A. et al. A multifaceted data mining approach to understanding what factors lead college students to persist and graduate. In: IEEE. *2017 Computing Conference*. [S.l.], 2017. p. 372–381. Citado 4 vezes nas páginas 34, 35, 36 e 37.

HÄMÄLÄINEN, W.; VINNI, M. Classifiers for educational data mining. *Handbook of Educational Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, p. 57–71, 2011. Citado 2 vezes nas páginas 21 e 25.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790. Citado 5 vezes nas páginas 20, 24, 25, 27 e 61.

HANS, R. T.; MNKANDLA, E. Modeling software engineering projects as a business: A business intelligence perspective. p. 1–5, Sep. 2013. ISSN 2153-0033. Citado 2 vezes nas páginas 17 e 19.

HARRIOTT, J. 7 pillars for successful analytics implementation. *Marketing Insights*, v. 25, n. 1, p. 34–41, 2013. Citado na página 19.

HE, H.; MA, Y. *Imbalanced learning: foundations, algorithms, and applications*. [S.l.]: John Wiley & Sons, 2013. Citado na página 29.

Hegde, V.; Prageeth, P. P. Higher education student dropout prediction and analysis through educational data mining. In: *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. [S.l.: s.n.], 2018. p. 694–699. Citado na página 42.

I., A. G. B. F.; ROSA, J. S. Sigaa mobile–o caso de sucesso da ferramenta de gestão acadêmica na era da computação móvel. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2013. v. 24, n. 1, p. 92. Citado 2 vezes nas páginas 51 e 69.

IHANTOLA, P. et al. Educational data mining and learning analytics in programming: Literature review and case studies. In: *Proceedings of the 2015 ITiCSE on Working Group Reports*. [S.l.: s.n.], 2015. p. 41–63. Citado 2 vezes nas páginas 34 e 36.

ISIK, O.; JONES, M. C.; SIDOROVA, A. Business intelligence success: The roles of bi capabilities and decision environments. *Information e Management*, v. 50, p. 13–23, 12 2012. Citado na página 19.

JEONG, S.; KIM, S.-W.; CHOI, B.-U. Dimensionality reduction in high-dimensional space for multimedia information retrieval. In: *Proceedings of the 18th International Conference on Database and Expert Systems Applications*. Berlin, Heidelberg: Springer-Verlag, 2007. (DEXA'07), p. 404–413. ISBN 3540744673. Citado na página 26.

JOHN, G. H.; KOHAVI, R.; PFLEGER, K. Irrelevant features and the subset selection problem. In: *MACHINE LEARNING: PROCEEDINGS OF THE ELEVENTH INTERNATIONAL*. [S.l.]: Morgan Kaufmann, 1994. p. 121–129. Citado na página 27.

JÚNIOR, J. G. d. O. et al. *Identificação de padrões para a análise da evasão em cursos de graduação usando mineração de dados educacionais*. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2015. Citado 2 vezes nas páginas 38 e 63.

KANTORSKI, G. et al. Predição da evasão em cursos de graduação em instituições públicas. v. 27, n. 1, p. 906, 2016. Citado na página 38.

KHAN, R.; QUADRI, S. M. K. Business intelligence: An integrated approach. 01 2019. Citado na página 12.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, v. 97, n. 1, p. 273 – 324, 1997. ISSN 0004-3702. Relevance. Disponível em: <http://www.sciencedirect.com/science/article/pii/S000437029700043X>. Citado 2 vezes nas páginas 27 e 28.

KOSTOPOULOS, G.; KOTSIANTIS, S.; PINTELAS, P. Estimating student dropout in distance higher education using semi-supervised techniques. In: *Proceedings of the 19th Panhellenic Conference on Informatics*. [S.l.: s.n.], 2015. p. 38–43. Citado 4 vezes nas páginas 12, 34, 36 e 37.

KOSTOPOULOS, G. et al. Early dropout prediction in distance higher education using active learning. In: IEEE. *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*. [S.l.], 2017. p. 1–6. Citado 3 vezes nas páginas 34, 36 e 37.

LAKKARAJU, H. et al. A machine learning framework to identify students at risk of adverse academic outcomes. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2015. p. 1909–1918. Citado 3 vezes nas páginas 34, 35 e 36.

LAM-ON, N.; BOONGOEN, T. Using cluster ensemble to improve classification of student dropout in thai university. p. 452–457, Dec 2014. Citado na página 40.

LAUDON, K.; LAUDON, E. *JP (2007) Sistemas de Informação Gerenciais. 7a. Edição*. [S.l.]: São Paulo: Prentice Hall, 2007. Citado na página 20.

LAURÍA, E. J. et al. Mining academic data to improve college student retention: An open source perspective. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. [S.l.: s.n.], 2012. p. 139–142. Citado 3 vezes nas páginas 34, 36 e 37.

LIANG, J.; LI, C.; ZHENG, L. Machine learning application in moocs: Dropout prediction. In: IEEE. *2016 11th International Conference on Computer Science & Education (ICCSE)*. [S.l.], 2016. p. 52–57. Citado 5 vezes nas páginas 12, 34, 35, 36 e 37.

LIMA, A.; JÚNIOR, M.; NASCIMENTO, A. Um survey com empresas brasileiras acerca da utilização de business intelligence (bi) e um diagnóstico sobre a infraestrutura e metodologias associadas; a survey with brazilian companies on the business intelligence (bi) use and a diagnosis on the infrastructure and associated methodologies. In: . [S.l.: s.n.], 2017. Citado 2 vezes nas páginas 12 e 20.

LIMA, A. S. de. *Proposta e avaliação da combinação de uma metodologia Ágil e gqm+strategies para o desenvolvimento de aplicações de business intelligence dirigido à estratégia*. Dissertação (Mestrado) — Universidade Federal de Sergipe, 2017. Citado 2 vezes nas páginas 17 e 18.

LIU, K. F.-R.; CHEN, J.-S. Prediction and assessment of student learning outcomes in structural mechanics a decision support of integrating data mining and fuzzy logic. In: IEEE. *2010 2nd International Conference on Education Technology and Computer*. [S.l.], 2010. v. 3, p. V3–499. Citado 2 vezes nas páginas 34 e 36.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. 2012. Citado na página 12.

MACHADO, R. D. et al. Estudo bibliométrico em mineração de dados e evasão escolar. In: *Congresso Nacional de Excelencia em Gestao*. [S.l.: s.n.], 2015. v. 8. Citado 2 vezes nas páginas 43 e 50.

MANHAES, L. et al. Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. 01 2012. Citado 3 vezes nas páginas 10, 13 e 38.

MANHÃES, L. M. B.; CRUZ, S. M. S. da; ZIMBRÃO, G. Wave: an architecture for predicting dropout in undergraduate courses using edm. In: *Proceedings of the 29th annual acm symposium on applied computing*. [S.l.: s.n.], 2014. p. 243–247. Citado 5 vezes nas páginas 11, 34, 35, 36 e 37.

MARCH, S. T.; HEVNER, A. R. Integrated decision support systems: A data warehousing perspective. *Decision Support Systems*, v. 43, p. 1031–1043, 2007. Citado na página 19.

MARQUEZ-VERA, C.; MORALES, C. R.; SOTO, S. V. Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, v. 8, n. 1, p. 7–14, Feb 2013. ISSN 1932-8540. Citado 6 vezes nas páginas 13, 34, 36, 37, 39 e 61.

MARTINS, L. C. B. et al. Early prediction of college attrition using data mining. In: IEEE. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.], 2017. p. 1075–1078. Citado 6 vezes nas páginas 11, 33, 34, 36, 37 e 43.

MEC, B. Documento orientador para a superação da evasão e retenção na rede federal de educação profissional, científica e tecnológica. 2014. Citado na página 11.

Medina, E. C. et al. Predictive model to reduce the dropout rate of university students in perú: Bayesian networks vs. decision trees. In: *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*. [S.l.: s.n.], 2020. p. 1–7. Citado na página 41.

MUSSO, M. F. et al. Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks. *Frontline Learning Research*, ERIC, v. 1, n. 1, p. 42–71, 2013. Citado na página 40.

MáRQUEZ, C. et al. Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, v. 33, p. 107–124, 02 2016. Citado 2 vezes nas páginas 11 e 62.

NASCIMENTO, R. F. F. et al. O algoritmo support vector machines (svm): avaliação da separação ótima de classes em imagens ccd-cbers-2. *Simpósio Brasileiro de Sensoriamento Remoto*, v. 14, p. 2079–2086, 2009. Citado na página 21.

NETO, A. de O. S. *Um sistema de BI de código aberto para o apoio às coordenações de curso do departamento de computação.* [S.l.], 2016. Citado 4 vezes nas páginas 44, 45, 88 e 89.

NURHUDA, A.; ROSITA, D. Prediction student graduation on time using artificial neural network on data mining students stmik widya cipta dharma samarinda. In: *Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government*. [S.l.: s.n.], 2017. p. 86–89. Citado 3 vezes nas páginas 34, 36 e 37.

OSMANBEGOVIC, E.; SULJIC, M. Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, Tuzla: University of Tuzla, Faculty of Economics, v. 10, n. 1, p. 3–12, 2012. Citado na página 41.

PADMAPRIYA, B.; VELMURUGAN, T. A survey on breast cancer analysis using data mining techniques. p. 1–4, 2014. Citado na página 21.

PANDA, S. K.; NAG, S.; JANA, P. K. A smoothing based task scheduling algorithm for heterogeneous multi-cloud environment. In: IEEE. *2014 International Conference on Parallel, Distributed and Grid Computing*. [S.l.], 2014. p. 62–67. Citado na página 28.

Perez, B.; Castellanos, C.; Correal, D. Applying data mining techniques to predict student dropout: A case study. In: *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*. [S.l.: s.n.], 2018. p. 1–6. Citado na página 42.

PRATI, R. C. et al. Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, v. 6, n. 2, p. 215–222, 2008. Citado na página 26.

PRISTYANTO, Y.; SETIAWAN, N. A.; ARDIYANTO, I. Hybrid resampling to handle imbalanced class on classification of student performance in classroom. In: IEEE. *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*. [S.l.], 2017. p. 207–212. Citado 3 vezes nas páginas 34, 36 e 37.

PUARUNGROJ, W. et al. Application of data mining techniques for predicting student success in english exit exam. In: *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication*. [S.l.: s.n.], 2018. p. 1–6. Citado 4 vezes nas páginas 11, 34, 36 e 37.

RAGAB, A. H. M. et al. A comparative analysis of classification algorithms for students college enrollment approval using data mining. In: *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments*. [S.l.: s.n.], 2014. p. 106–113. Citado 3 vezes nas páginas 34, 36 e 37.

RODRIGUES, M. W.; ISOTANI, S.; ZÁRATE, L. E. Educational data mining: A review of evaluation process in the e-learning. *Telematics and Informatics*, Elsevier, v. 35, n. 6, p. 1701–1717, 2018. Citado 2 vezes nas páginas 34 e 36.

ROMERO, C.; VENTURA, S. Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 7, n. 1, p. e1187, 2017. Citado 2 vezes nas páginas 34 e 36.

ROMERO, C. et al. Data mining algorithms to classify students. In: *Educational data mining 2008*. [S.l.: s.n.], 2008. Citado na página 21.

ROY, S.; GARG, A. Analyzing performance of students by using data mining techniques a literature survey. In: IEEE. *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*. [S.l.], 2017. p. 130–133. Citado 3 vezes nas páginas 23, 34 e 36.

ROY, S.; GARG, A. Predicting academic performance of student using classification techniques. In: IEEE. *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*. [S.l.], 2017. p. 568–572. Citado na página 40.

RUSTIA, R. A. et al. Predicting student's board examination performance using classification algorithms. In: *Proceedings of the 2018 7th International Conference on Software and Computer Applications*. [S.l.: s.n.], 2018. p. 233–237. Citado 3 vezes nas páginas 34, 36 e 37.

SAA, A. A. Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications*, v. 7, n. 5, p. 212–220, 2016. Citado na página 23.

SANTOS, J. S. d. Business intelligence: uma proposta metodológica para análise da evasão escolar em instituições federais de ensino. 2017. Citado na página 12.

SANTOS, M. d. A. e. a. Aplicação de algoritmos de árvore de decisão na previsão da evasão escolar: um estudo no campus lagarto do ifs. 2017. Citado na página 10.

SHAUN, R. et al. Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19, p. 3–13, 01 2011. Citado na página 37.

SILVA, L. d. Mineração de dados: uma abordagem introdutória e ilustrada. *Editora Mackenzie (Coleção conexão inicial, v. 11)*, v. 1, 2015. Citado na página 69.

SLIM, A. et al. Predicting student success based on prior performance. In: IEEE. *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. [S.l.], 2014. p. 410–415. Citado 3 vezes nas páginas 34, 35 e 36.

SPACKMAN, K. A. Signal detection theory: Valuable tools for evaluating inductive learning. In: ELSEVIER. *Proceedings of the sixth international workshop on Machine learning*. [S.l.], 1989. p. 160–163. Citado na página 26.

SUKHIJA, K.; JINDAL, M.; AGGARWAL, N. The recent state of educational data mining: A survey and future visions. In: IEEE. *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*. [S.l.], 2015. p. 354–359. Citado 3 vezes nas páginas 23, 34 e 36.

TEIXEIRA, I. *Instituto Nacional de Estudos e P. E. A. Censo da educação superior 2015*. [S.l.], 2015. Citado na página 10.

TURBAN, E. et al. *Business intelligence: a managerial perspective on analytics*. [S.l.]: Prentice Hall, New York, 2013. Citado na página 20.

VASCONCELOS, N. O. *Data Mining e Data Analytics para Apoio à Gestão Estratégica e Mitigação da Evasão Escolar*. Dissertação (Mestrado) — Universidade Federal de Sergipe, 2019. Citado na página 41.

VERIKAS, A. et al. Electromyographic patterns during golf swing: Activation sequence profiling and prediction of shot effectiveness. *Sensors*, v. 16, p. 592, 04 2016. Citado na página 22.

VERLEYSEN, M.; FRANCOIS, D. The curse of dimensionality in data mining and time series prediction. In: *Proceedings of the 8th International Conference on Artificial Neural Networks: Computational Intelligence and Bioinspired Systems*. Berlin, Heidelberg: Springer-Verlag, 2005. (IWANN'05), p. 758–770. ISBN 3540262083. Disponível em: <https://doi.org/10.1007/11494669_93>. Citado na página 26.

WARR, W. A. Scientific workflow systems: Pipeline pilot and knime. *Journal of computer-aided molecular design*, Springer, v. 26, n. 7, p. 801–804, 2012. Citado na página 29.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123748569. Citado 2 vezes nas páginas 27 e 28.

YADAV, S.; SHUKLA, S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: IEEE. *2016 IEEE 6th International conference on advanced computing (IACC)*. [S.l.], 2016. p. 78–83. Citado na página 26.

YOO, I. et al. Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, Springer Nature, v. 36, n. 4, p. 2431–2448, may 2011. Disponível em: <https://doi.org/10.1007%2Fs10916-011-9710-5>. Citado na página 11.